January 2011

Predictive Alternatives in Bayesian Model Selection

Andrew Womack
Washington University in St. Louis

WASHINGTON UNIVERSITY

Department of Mathematics

Dissertation Examination Committee:

Jefferson Gill, Chair

Siddhartha Chib

Edward Greenberg

Nan Lin

Edward Spitznagel

Mladen Victor Wickerhauser

PREDICTIVE ALTERNATIVES IN BAYESIAN MODEL SELECTION

by

Andrew Womack

A dissertation presented to the
Graduate School of Arts and Sciences
of Washington University in
partial fulfillment of the
requirements for the degree
of Doctor of Philosophy

May 2011

St. Louis, Missouri

ABSTRACT

Predictive Alternatives in Bayesian Model Selection

by

Womack, Andrew

Doctor of Philosophy in Mathematics,

Washington University in St. Louis, May, 2011.

Professor Jeff Gill, Chairperson

Model comparison and hypothesis testing is an integral part of all data analyses. In this thesis, I present two new families of information criteria that can be used to perform model comparison. In Chapter 1, I review the necessary background to motivate the thesis. Of particular interest is the role of priors for estimation and model comparison as well as the role that information theory can play in the latter. As we will see, many existing forms of model comparison can be viewed in an information theoretic manner, which motivates defining new families of criteria. In Chapter 2, I present the two new criteria and discuss their properties. The first criterion is based purely on posterior predictive densities and Kullback-Leibler divergences and decomposes into terms that describe the fit and complexity of the model. In this manner, it behaves similar to popular criteria, such as the AIC or the DIC. I then present

the second family of criteria, which are a modification of the marginal distribution by an appropriate Rényi divergence. This modification of the marginal allows the investigator to use priors that reflect vague prior knowledge while not suffering the paradoxes that can arise from such priors. One particularly nice aspect of this family of criteria is that it subsumes the Bayes' factor as a special case and produces an infinite family of criteria that are asymptotically equivalent to the Bayes' factor. In this manner, the criteria can be modified to achieve certain goals in small samples while maintaining asymptotic consistency. I conclude the thesis with a short discussion of the computational difficulties that arise when using the criteria and explore possible ways to overcome them.

ACKNOWLEDGMENTS

As I reflect on the past six years spent at Washington University in St. Louis, it is clear to me that there are many people to whom I am indebted for their ceaseless support. Most obviously, this dissertation project benefited enormously from the support of my committee members, especially my Chair, Dr. Jeff Gill. Jeff provided me with countless constructive conversations and carefully read every draft of my dissertation. Through the Center for Applied Statistics, he provided me with the resources and education to complete my dissertation project. In addition, I would like to thank Jeff for the opportunity to serve as Teaching Fellow for two Distinguished Visiting Statistician Courses through which I learned a great deal and made connections outside of Washington University. I would like to thank Sid Chib for the inspiration that his career has offered me and his devotion to subjective Bayesian statistics, especially in helping me to understand its power in answering important questions in the world. I would like to thank Ed Greenberg for his enthusiasm in joining my committee shortly before my defense as well as for providing a number of insightful questions that will help me move my research into future projects. I would like to thank Nan Lin for giving me the academic freedom to pursue research the research that I wanted to pursue, as well as for numerous insightful suggestions over the years. I would like to thank Victor Wickerhauser for introducing to many topics in applied mathematics

and helping me to understand the merits and pitfalls of various approaches to these problems. I would like to thank Ed Spitznagel for reigniting my love of statistics when I TAed his course as well as his friendship and humor, which made many days much more pleasant than they otherwise would have been. In addition to thanking my committee, I would like to thank the faculty of the Department of Mathematics. Without the great instruction of the faculty members, I would not have achieved the mathematical maturity necessary to complete my PhD. In particular, I would like to thank Guido Weiss and Stanley Sawyer for their excellent instruction in analysis and probability theory. Finally, I would like to thank my fellow graduate students (both in Mathematics and other departments) for their support and friendship over the years.

Above all else, I thank my family for there love and support throughout the years. Jim, Kathlleen, Katie, Charlie, Jen, Mary, and Kassidie have been integral in giving me the resolve to finish my graduate education. From my earliest days in primary school through these last few years, their confidence in me has never wavered and has provided me with the strength to continue. Additionally, I would like to thank my fiancée Diana O'Brien for her infinite love and patience as well as her family (John, Anna, and Valerie) for their support. Finally, I would like to thank my fellow students for friendship, in particular: Sean Mueller, Ian Ostrander, Scott Cook, and Sara Gharabeigi.

This thesis is dedicated to Diana O'Brien, without whose support it could not have been completed.

TABLE OF CONTENTS

Page

ABSTRACT ii

ACKNOWLEDGMENTS iv

1 **Background** 1

1.1 Estimation 1

1.1.1 Bayes Theorem 1

1.1.2 Exchangeability and Priors 2

1.1.3 Posterior Consistency 5

1.1.4 The Roles of Priors 6

1.1.5 Prior Specification 8

1.1.6 Improper Priors 10

1.2 Model Comparison 14

1.2.1 Bayes Factors 14

1.2.2 A Paradox from Improper Priors 16

1.2.3 Overcoming Issues with Improper Priors 18

1.2.4 Model Focus and Selection Criteria 20

Page

1.3 Information Theory 24

1.3.1 Shannon Entropy and K-L Divergence 25

1.3.2 Rényi Divergences 28

2 **New Alternatives for Model Selection** 33

2.1 Posterior Predictive Information Criterion 33

2.2 The PPIC 38

2.2.1 Examples 40

2.2.2 Multiple Model Comparison 46

2.3 Predictively Modified Bayes' Factors 48

2.3.1 Iwaki's Expected Posterior Predicted Priors 50

2.3.2 General Properties 53

2.3.3 Choice of α 59

2.3.4 Analytical Examples 62

2.3.5 Computational Examples 67

2.4 Computational Issues 76

2.5 Conclusions 78

1. Background

This chapter presents the basic constructions and results from Bayesian statistics that motivate the thesis. In order to place the thesis in the proper context, we consider traditional Bayesian problems of estimation and selection. In particular, we consider parametric models with parameters $\theta_k \in \Theta_k$ where Θ_k is a p_k dimensional space, taken as a subset of $\mathbb{R}^k$. Within this class of models, we provide the basic background on estimation, prediction, and model comparison.

1.1 Estimation

1.1.1 Bayes Theorem

There are two basic tenets of Bayesian statistics: (1) all unknown values are given probability distributions and (2) beliefs are updated via Bayes' rule

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}, \tag{1.1}$$

where $P(A|B)$ is the conditional probability that event A occurs if event B has already occurred. In the context of statistics, one usually takes A to be some quantity of interest and B the observed data. There are several alternative approaches to motivating and defining probability theory and formula (1.1). Following the work of

Kolmogorov [1, 2], a measure theoretic approach views (1.1) as a consequence of the measure theoretic definition of conditional probability. In contrast, statisticians such as Rényi [3] and Cox [4] place conditional probabilities in a more central role. Cox's framework–as discussed in Jaynes' [5] presents some desirable properties for extending Aristotelian logic from deductive to inductive logic. Probability theory (at least with finite additivity) is derived as the unique system of logic to achieve the desired goal. The essential elements of the theorem are: the positive real representation for plausibilities,; the functional relationship between the plausibility of a statement and the plausibility of its negation; associativity in reasoning about conjunctions and a density argument [6]. The conclusion of the theorem is that the product rule $P(A \cap B) = P(A|B)P(B) = P(B|A)P(A)$ and the complement rule $P(A^C) = 1 - P(A)$ are the unique rules needed to extend Aristotelian logic and the rules for disjunctions are derived as a consequence of the rules for conjunctions and negations. While this theorem provides a natural motivation of tenet (2), it does not justify tenet (1).

1.1.2 Exchangeability and Priors

The most complete way to motivate tenet (1) is to invoke theorems about exchangeable random variables [7, 8]A sequence of random variables $X_1, \ldots, X_n, \ldots$ is said to be exchangeable if the distribution of $X_1, \ldots, X_k$ is the same as the distribution of $X_{\tau(1)}, \ldots, X_{\tau(k)}$ for all premutations τ in the group $\mathcal{S}_k$ and for all $k \in \mathbb{N}$. The deFinetti-Hewitt-Savage theorem states that a sequence of exchangeable random

variables can be represented by a sequence of conditionally independent and identically distributed random variables, when conditioning on a particular measure μ on the space of probability measures. In particular, if $X_1, \ldots, X_n, \ldots$ is an exchangeable sequence, then

$$\begin{aligned} P(X_1 \in A_1, \ldots, X_n \in A_n) &= \int P(X_1 \in A_1, \ldots, X_n \in A_n | Q) \mathrm{d}\mu(Q) \\ &= \int P(X_1 \in A_1 | Q) \cdots P(X_n \in A_n | Q) \mathrm{d}\mu(Q) \\ &= \int Q(A_1) \cdots Q(A_n) \mathrm{d}\mu(Q). \end{aligned} \tag{1.2}$$

In terms of a generative process for the data, one assumes that Q is chosen via μ before any data are generated and the X_i are then generated by that Q. The simplest Bayesian models are formed when one considers an exchangeable sequence of random variables $\{X_i \ : \ i = 1, \ldots, n\}$, with realizations $\{x_i \ : \ i = 1, \ldots, n\}$, which are assumed to arise as iid draws from a distribution $Q(A) = F(A|\boldsymbol{\theta})$. The mixing measure is taken as $\Pi(B) = P(\boldsymbol{\theta} \in B)$. Essentially, one makes the subjective determination that the distribution of Q is restricted to a finite dimensional subspace of the space of all probability measures. The $F(A|\boldsymbol{\theta})$ proposed is the likelihood function and $\Pi(B)$ is the prior distribution of $\boldsymbol{\theta}$. These are usually assumed to be absolutely continuous with respect to some measures ($\mathrm{d}x$ and $\mathrm{d}\boldsymbol{\theta}$) and the densities are given by $f(x|\boldsymbol{\theta})$ and $\pi(\boldsymbol{\theta})$. One then uses Bayes' rule to update beliefs about $\boldsymbol{\theta}$ by computing the posterior density of $\boldsymbol{\theta}$,

$$p(\boldsymbol{\theta}|x_1, \ldots, x_n) = \frac{f(x_1, \ldots, x_n|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{m(x_1, \ldots, x_n)}. \tag{1.3}$$

The normalizing constant $m(x_1, \ldots, x_n)$ is given by the marginal density of the x_i,

$$m(x_1, \ldots, x_n) = \int f(x_1, \ldots, x_n|\boldsymbol{\theta})\pi(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}. \tag{1.4}$$

As we will see, this marginal is key to Bayesian model comparison and can be sensitive to misrepresentations of the prior $\pi(\boldsymbol{\theta})$. When performing a statistical analysis, if we had access to π, then the problem can be trivial. Even given an assumption like $\boldsymbol{X}$ are exchangeable from a particular distribution, the theorem can provide restrictions on the likelihood $f(x|\boldsymbol{\theta})$, but offers no insight into the class of π which could generate the data. de Finetti originally proved this theorem for the case where the $\boldsymbol{X}_i$ are exchangeable Bernoulli random variables. The likelihood is a Bernoulli distribution with parameter $0 < \theta < 1$ and the prior is some mixing measure which is unrestricted on $(0, 1)$. The fact that the assumption of exchangeability provides no restriction for the mixing measure even in this simple case provides quite a strong argument for the Subjective Bayesian interpretation of probability put forth by De Finetti, Savage, and Lindley. We were able to make assumptions to limit the likelihood, but any inferences about marginal distributions depend upon choices for a mixing distribution and these are subject to the personal beliefs and predilections of the investigator (or, perhaps less personally, they are subject to the beliefs of a group or population of investigators).

1.1.3 Posterior Consistency

Though the D-H-S theorem provides no restriction on the prior π, when making inferences about $\boldsymbol{\theta}$ through Bayes' rule the influence of π can be asymptotically negligible. Given the generative process assumed for the data, each particular dataset is to be generated by a specific value of $\boldsymbol{\theta}$. Calling this value $\boldsymbol{\theta}_0$, we ask: "Does $p(\boldsymbol{\theta}|x_1, \ldots, x_n)$ concentrate around the true value $\boldsymbol{\theta}_0$ as $n \to \infty$?" The idea of concentration is given by the notion of consistency [9]: a posterior is said to be consistent at $\boldsymbol{\theta}_0$ if there exists a set of sequences $\boldsymbol{X}_n(\omega)$–indexed by ω–with probability 1 (given $\boldsymbol{\theta} = \boldsymbol{\theta}_0$) such that for all neighborhoods U of $\boldsymbol{\theta}_0$ we have

$$\Pi(U|\boldsymbol{X}_n(\omega)) \to 1.$$

What is most important for us is that two reasonable priors will eventually give rise to posteriors that are very similar. Thus, prior distributions that are mis-specified or are weak (i.e. diffuse) are eventually overcome by the data. However, one must take caution against pathological priors, such as those discussed

Theorem 1.1.1 (Posterior Robustness)**.** *Suppose that the likelihood function is $f(x|\boldsymbol{\theta})$ and that $\boldsymbol{\theta}_0$ is in the interior of Θ. Let π_1, π_2 be two priors that are positive and continuous at $\boldsymbol{\theta}_0$ such that their posteriors are consistent at $\boldsymbol{\theta}_0$. Then*

$$\lim_{n\to\infty} \int |\pi_1(\boldsymbol{\theta}|\boldsymbol{x}_n) - \pi_2(\boldsymbol{\theta}|\boldsymbol{x}_n)| d\boldsymbol{\theta} = 0$$

with probability 1.

Proof. (see [9]) □

1.1.4 The Roles of Priors

Beyond these simple statistical models, many models are given in a hierarchical fashion where partial, rather than full, exchangeability is assumed. In these models, the data often come in groupings $x_j = \{x_{1j}, \ldots, x_{n_j j}\}$ for $j = 1, \ldots, J$. The first exchangeability assumption means that both members within each group and the groups themselves are exchangeable. This gives rise to a parameter $mbf\phi$, with respect to which the groups are conditionally independent. It also gives rise to group level variables $\boldsymbol{\theta}_j$, with respect to which individuals within a group are conditionally independent,

$$x_{ij} \sim f(x|\boldsymbol{\theta}_j, \boldsymbol{\phi})$$

$$\boldsymbol{\theta}_j \sim \pi(\boldsymbol{\theta}|\boldsymbol{\phi})$$

$$\boldsymbol{\phi} \sim \pi(\boldsymbol{\phi}).$$

In this type of model, the prior distribution plays two roles. First, it induces the correlation structure of the data ($\pi(\boldsymbol{\theta}|\boldsymbol{\phi})$). In this manner, this prior is used in a modeling context, providing structure to the data and shrinking the $\boldsymbol{\theta}_j$ towards areas of high probability according to $\pi(\boldsymbol{\theta}|\boldsymbol{\phi})$. Second, the prior aids in estimation of ϕ

through $(\pi(\boldsymbol{\phi}))$. To see this more clearly, the parameters $\boldsymbol{\theta}_j$ can be viewed as a parameter expanded version of the likelihood function for the data,

$$\begin{aligned} f(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_J|\boldsymbol{\phi}) &= \prod_j f(\boldsymbol{x}_j|\boldsymbol{\phi}) \\ &= \prod_j \int f(\boldsymbol{x}_j|\boldsymbol{\theta}_j, \boldsymbol{\phi})\pi(\boldsymbol{\theta}_j|\boldsymbol{\phi})\mathrm{d}\boldsymbol{\theta}_j \\ &= \int \prod_j \left(f(\boldsymbol{x}_j|\boldsymbol{\theta}_j, \boldsymbol{\phi})\pi(\boldsymbol{\theta}_j|\boldsymbol{\phi})\right) \mathrm{d}\boldsymbol{\theta}_1 \ldots \mathrm{d}\boldsymbol{\theta}_j \\ &= \int \prod_{i,j} \left(f(x_{ij}|\boldsymbol{\theta}_j, \boldsymbol{\phi})\pi(\boldsymbol{\theta}_j|\boldsymbol{\phi})\right) \mathrm{d}\boldsymbol{\theta}_1 \ldots \mathrm{d}\boldsymbol{\theta}_j, \end{aligned}$$

where Fubini's rule has been applied after making the appropriate measurability assumptions. Alternatively, one could also interpret this model as though the $\boldsymbol{\theta}_j$ are the parameters that are of most interest and not knowing $\boldsymbol{\phi}$ is merely a nuisance. In this case, the likelihood of interest is

$$\begin{aligned} f(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_J|\boldsymbol{\theta_1}, \ldots, \boldsymbol{\theta}_j) &= \int f(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_J|\boldsymbol{\theta_1}, \ldots, \boldsymbol{\theta}_J, \boldsymbol{\phi})\pi(\boldsymbol{\phi}|\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_J)\mathrm{d}\boldsymbol{\phi} \\ &= \int f(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_J|\boldsymbol{\theta_1}, \ldots, \boldsymbol{\theta}_J, \boldsymbol{\phi})\frac{\pi(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_J|\boldsymbol{\phi})\pi(\boldsymbol{\phi})}{\pi(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_J)}\mathrm{d}\boldsymbol{\phi} \\ &= \frac{\int \prod_j \left(f(\boldsymbol{x}_j|\boldsymbol{\theta}_j, \boldsymbol{\phi})\pi(\boldsymbol{\theta}_j|\boldsymbol{\phi})\right)\pi(\boldsymbol{\phi})\mathrm{d}\boldsymbol{\phi}}{\pi(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_J)} \\ &= \frac{\int \prod_{ij} \left(f(x_{ij}|\boldsymbol{\theta}_j, \boldsymbol{\phi})\pi(\boldsymbol{\theta}_j|\boldsymbol{\phi})\right)\pi(\boldsymbol{\phi})\mathrm{d}\boldsymbol{\phi}}{\int \prod_j \left(\pi(\boldsymbol{\theta}_j|\boldsymbol{\phi})\right)\pi(\boldsymbol{\phi})\mathrm{d}\boldsymbol{\phi}}, \end{aligned}$$

where partial exchangeability has been used in the expression $\pi(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_J|\boldsymbol{\phi}) = \prod_j \pi(\boldsymbol{\theta}_j|\boldsymbol{\phi})$.

When a statistical problem calls for treating $\boldsymbol{\phi}$ as the true parameter that generates the data, then the first interpretation is used and a full probability model is

specified. However, if $\boldsymbol{\phi}$ is merely a nuisance then one can often provide an estimate of the posterior of interest $p(\boldsymbol{\theta_1}, \ldots, \boldsymbol{\theta}_j | \boldsymbol{x}_1, \ldots, \boldsymbol{x}_J)$ by using an Empirical Bayes approach. This approach treats $\boldsymbol{\phi}$ as though it is just a number and ignores any distributional assumptions made about it. One simply finds the value $\boldsymbol{\phi}^*$ which maximizes $f(\boldsymbol{x}|\boldsymbol{\eta})$ and inferences are drawn about the $\boldsymbol{\theta}_j$ using the conditional posterior density $p(\boldsymbol{\theta}_j|\boldsymbol{x}, \boldsymbol{\phi}^*)$. In this context, the parameter $\boldsymbol{\phi}$ can also be considered to be a tuning parameter and many researchers simply change the value of the parameter until the desired posterior behavior is observed for the problem.

1.1.5 Prior Specification

For this project, we are concerned with full probability specifications. When one implements a hierarchical model (and thus most of the prior specification is used for modeling heterogeneity in the data and the last part is used for estimation purposes) the problem often arises as to how one should specify the prior $\pi(\boldsymbol{\phi})$. There are many possible choices, but 1.1.1 suggests that as long as we maintain positivity, continuity, and consistency then the posteriors will eventually become very "similar" . There are essentially three alternative approaches to specifying the prior. The first is to fully specify a subjective prior for $\boldsymbol{\phi}$ [10, 11], the second is to choose a convenient family of conditionally conjugate priors (which will aid in MCMC estimation), and the third is to use some default "objective" prior . The first method is considered ideal as it provides a full probability analysis which represents the expert beliefs of

the practitioner or those that have been agreed upon in a field. Nonetheless, these priors often present computing challenges, are difficult to elicit, and must be elicited for each model. Alternatively, the second method can provide a good approximation to a full probability analysis while presenting fewer challenges in computation and elicitation (though they still require an elicitation for each model). Finally, the third option is advantageous when there are many models to consider (and so full or partial elicitation is too time consuming) or there is only weak prior knowledge (for example that $\boldsymbol{\phi}_0$ lies within some set). In these situations, the full probability model is not specified; rather, one can use some particular rules for defining priors that represent a minimal amount of information. There are several issues with this final method. First, if diffuse priors were used, one must be sure that these correspond to integrable (proper) posteriors for some minimal sample size . Second, it is often hard to determine what "diffuse" means in different contexts. Third, as will be shown, these priors are problematic for model comparison. Despite these caveats, these priors often provide "nice" posterior distributions (i.e. frequentist matching, unbiased). Moreover, practitioners may prefer "objective" priors because they have weak prior knowledge or due to the difficulties associated with prior elicitation and having only a finite amount of resources devoted to gaining knowledge from experts that might be better spent in building likelihoods. [12]

Before the discussion of various default "objective" priors, we present a brief example of a simple model that exhibits consistency even for prior distributions that

become increasing diffuse. We present this example for three reasons. First, the conjugate framework is easy to see. Second, the entire class of conjugate priors provides consistency. Third points on the boundary of set of hyperparameters will also provide consistent posteriors. However, as we will see later, letting the hyperparameters tend towards the boundary is problematic when comparing models.

Example 1.1.1 (Exponential Distribution)**.** *Suppose that* $x_1, \ldots, x_n$ *are iid* $\mathcal{E}(\theta)$ *and* $\theta \sim$ *Gamma*(α, β) *for some* $\alpha, \beta > 0$*. Then the posterior* $p(\theta|\boldsymbol{x}_n, \alpha, \beta)$ *is Gamma*$(\alpha + n, \beta + \sum x_i)$*. For any given* (α, β) *which are positive and finite, these posteriors converge weakly to* $\delta_{\theta_0}(\theta)$ *where* $\theta_0 = \lim_{n\to\infty} \frac{n}{\sum_i x_i}$*. In fact, letting* $(\alpha, \beta) = (0, 0)$ *provides the prior* $\pi(\theta) \propto \theta^{-1}$ *and posterior Gamma*$(n, \sum x_i)$*, which also converges weakly to* $\delta_{\theta_0}(\theta)$*. This prior provides a posterior which is exactly frequentist matching*[1] *and provides unbiased estimation of* θ_0 *via the posterior mean.*

1.1.6 Improper Priors

One of Fisher's primary objections to the use of Bayes' rule in statistical estimation is that definitions of priors are not necessarily invariant to transformations of the parameter [13] . While it is true that a prior produced under a specific parameterization transforms by the appropriate rules, how a prior is built depends on how one reasons in particular parameterizations. A simple example is reasoning about a parameter $\theta \in [0, 1]$. Since $[0, 1]$ is compact, it might be reasonable to use a uniform prior for

[1]The posterior distribution of the parameter and the sampling distribution of an UMVUE coincide.

θ. However, the same reasoning applies to any reparameterization $\theta \mapsto \theta^a$ for any $a > 0$. One can therefore easily object to using uniform priors on reasonable compact sets. Though Fisher's objection does not apply to any particular prior (which can be defended by its creator), it does cast doubt on the practice of defining methods through which one should define priors. For the elicitation of a proper prior, one often finds an easy to understand parameterization and provides some structure to the prior (conjugate family for example). The hyperparameters are then changed until the prior reflects expert knowledge. This method, though convenient in practice, has a lot of subjective choices involved and making different choices could heavily influence the prior derived.

One approach is to devise a method to define priors that are invariant to certain transformations. Jefferys devised the first two approaches along these lines. The most obvious way to define priors which have an invariance property is to use a group G of transformations of $\boldsymbol{\theta}$ and seek a measure which is invariant to the left (or right) group action. This measure, denoted by π^G is the left (right) Haar measure associated to the parameter space [5,14] . Often the space of transformations is not compact and so the measure π^G need not be finite. In these cases, one obtains what is called an improper prior because it lacks integrability and therefore cannot represent subjective beliefs. These priors are dependent upon finding an appropriate class of transformations and the choice of this set of transformation can often be difficult. However, can sometimes

be induced through an allowable set of transformation of the data. A simple example is a location-scale family, presented presently.

Example 1.1.2 (Invariant Prior for Location-Scale Family)**.** *Consider* $f(x|\mu,\sigma) = \frac{1}{\sigma} f\left(\frac{x-\mu}{\sigma}\right)$ *where* f *represents the error structure of some physical process, for example measuring a distance with a ruler. An appropriate class of transformations of the data allows the investigator to change the point of initial measure and the units used for the measurement. The transformation is* $x' = \alpha(x - x_0)$*. To see that the likelihood is invariant to such transformations, note that*

$$\frac{1}{\sigma} f\left(\frac{x-\mu}{\sigma}\right) dx = \frac{1}{\sigma} f\left(\frac{\alpha((x-x_0)-(\mu-x_0))}{\alpha\sigma}\right) dx = \frac{1}{\sigma'} f\left(\frac{x'-\mu'}{\sigma'}\right) dx'$$

where $\mu' = \alpha(\mu - x_0)$ *and* $\sigma' = \alpha\sigma$*. If we want to induce this same structure on the prior, then*

$$\pi(\mu,\sigma)\, d\mu\, d\sigma = \pi(\mu',\sigma')\, d\mu'\, d\sigma' = \alpha^2 \pi(\alpha(\mu - x_0), \alpha\sigma)\, d\mu\, d\sigma$$

The general solution of this equation is $\pi(\mu,\sigma) \propto \sigma^{-2}$*. It is interesting to note that this prior also corresponds to the prior under Jeffreys' more general rule.*

When a specific set of transformations cannot be found, Jeffreys provided the general rule of defining $\pi^J(\boldsymbol{\theta}) \propto \sqrt{|I(\boldsymbol{\theta})|}$ where $I(\boldsymbol{\theta})$ is the Fisher information with $(I(\boldsymbol{\theta}))_{ij} = Cov(\partial_{\theta_i} \log(f(\boldsymbol{x}|\boldsymbol{\theta})), \partial_{\theta_j} \log(f(\boldsymbol{x}|\boldsymbol{\theta})))$ [15] . It is easy to see that this definition of the prior is invariant to both transformations of $\boldsymbol{\theta}$ and to transformations of $\boldsymbol{x}$, although it does depend upon $\boldsymbol{x}$ in some manner (for instance if the x_i are iid given $\boldsymbol{\theta}$, then the Jeffreys' prior is multiplied by the number of observations). When

these priors are improper, they are considered as defined only up to a multiplicative constant. However, there is at least one case where the prior obtained is in fact proper.

Example 1.1.3 (Jeffreys' Prior for Bernoulli Data)**.** *Suppose that the data are iid Bernoulli with parameter* θ*. The Fisher information for a single observation is* $\theta^{-1}(1-\theta)^{-1}$ *and so the Jeffrey's prior is* $\pi^J(\theta) \propto \theta^{-.5}(1-\theta)^{-.5}$ *which can be normalized and gives rise to a* $Beta(.5, .5)$ *prior distribution.*

Jeffreys' reasoning has been extended (see [16, 17]to the definition of reference priors (π^N), which seek to minimize the expected mutual information between the posterior and prior distributions,

$$I = \int \int \log \left(\frac{p(\boldsymbol{\theta}|\boldsymbol{x}_n)}{\pi(\boldsymbol{\theta})} \right) p(\boldsymbol{\theta}|\boldsymbol{x}_n) \mathrm{d}\boldsymbol{\theta} m(\boldsymbol{x}_n) \mathrm{d}\boldsymbol{x}_n \tag{1.5}$$

in the case when $n \to \infty$ and *supp* π is constrained to a sequence of nested compact sets whose union is the entire parameter space. Inherent in the reference prior definition is the order of the parameters in the model, finding priors first for those that are deemed the most important and moving through the parameter list sequentially. When this is ignored, one simply recovers the Jeffreys' prior. When ordering is used, the priors obtained often behave better in multidimensional settings than the Jeffreys' prior (and often correspond to a modified or independence Jeffreys' prior). The advantage of the reference prior definition over the Jeffreys' definition is that the Fisher information often does not exist but the mutual information does. When these

default priors give rise to proper posteriors, they can be used to perform a default data analysis. In addition to being used as priors, these measures can also be used as base measures when defining minimum information priors subject to prior constraints. Defining $A = \int \log\left(\frac{\pi(\boldsymbol{\theta})}{\pi^N(\boldsymbol{\theta})}\right)\pi(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}$, a few simple constraints can be elicited from an expert (e.g. support of moments) and the proper prior π, which minimizes A subject to those constraints, can be found. [18].

1.2 Model Comparison

1.2.1 Bayes Factors

In addition to the important goal of estimating the parameters of a model, statistics is also concerned with different forms of hypothesis testing. In the Bayesian framework, this is achieved through the use of probability theory. We have a finite number of models $M_k \ : \ k = 1, \ldots, K$, each equipped with a sampling density $f_k(\boldsymbol{x}|\boldsymbol{\theta}_k)$ and prior $\pi_k(\boldsymbol{\theta}_k)$. The models themselves also have prior probabilities $\pi(M_k)$. If we define $\boldsymbol{\theta} = \{\boldsymbol{\theta}_k \ : \ k = 1, \ldots, K\}$, $f(\boldsymbol{x}|\boldsymbol{\theta}, M_k) = f_k(\boldsymbol{x}|\boldsymbol{\theta}_k)$, and $\pi(\boldsymbol{\theta}|M_k) = \pi_k(\boldsymbol{\theta}_k)$, then

$$p(M_k|\boldsymbol{x}) = \frac{m(\boldsymbol{x}|M_k)\pi(M_k)}{m(x)} \tag{1.6}$$

where

$$m(\boldsymbol{x}|M_k) = \int f(\boldsymbol{x}|\boldsymbol{\theta}, M_k)\pi(\boldsymbol{\theta}|M_k)\mathrm{d}\boldsymbol{\theta} = \int f_k(\boldsymbol{x}|\boldsymbol{\theta}_k)\pi_k(\boldsymbol{\theta}_k)\mathrm{d}\boldsymbol{\theta} = m_k(\boldsymbol{x})$$

$$m(x) = \sum_{k=1}^{K} m(\boldsymbol{x}|M_k)\pi(M_k).$$

Defining the Bayes' Factor [19] as $BF_{\ell k}(\boldsymbol{x}) = \frac{m_\ell(\boldsymbol{x})}{m_k(\boldsymbol{x})}$, it is easy to see that the ratio of probabilities of two models is $\frac{p(M_\ell|\boldsymbol{x})}{p(M_k|\boldsymbol{x})} = BF_{\ell k}(\boldsymbol{x})\frac{\pi(M_\ell)}{\pi(M_k)}$ and that (1.6) can be re-written as

$$p(M_k|\boldsymbol{x}) = \left(\sum_{\ell=1}^{K} \frac{p(M_\ell|\boldsymbol{x})}{p(M_k|\boldsymbol{x})}\right)^{-1} = \left(\sum_{\ell=1}^{K} BF_{\ell k}(\boldsymbol{x})\frac{\pi(M_\ell)}{\pi(M_k)}\right)^{-1}.$$

An important result concerning Bayes factors comes from the notion of the merging of predictive densities. Given two priors for the same model that give rise to consistent posteriors, the predictive densities $pr_\ell(\boldsymbol{z}|\boldsymbol{x}_n) = \frac{m_\ell(\boldsymbol{z},\boldsymbol{x}_n)}{m_\ell(\boldsymbol{x}_n)}$ converge weakly to the same distribution almost surely (given $\boldsymbol{\theta}_0$) [9]. We can use this result to show that two marginals for the same model under different priors provide similar asymptotic behavior. Take the sequence of data $\boldsymbol{x}$ and separate it into two pieces $\boldsymbol{x}_n = \{x_1, \ldots, x_n\}$ and $\boldsymbol{z}_n = \{x_{n+1}, \ldots\}$. We can find n large enough such that the difference

$$\left| m_\ell(\boldsymbol{x}) - m_k(\boldsymbol{x})\frac{m_\ell(\boldsymbol{x}_n)}{m_k(\boldsymbol{x}_n)} \right| = |m_\ell(\boldsymbol{z}_n|\boldsymbol{x}_n) - m_k(\boldsymbol{z}_n|\boldsymbol{x}_n)|m_\ell(\boldsymbol{x}_n)$$

can be made arbitrarily small so that the two marginal distributions agree up to a multiplicative constant. This property of the marginal distributions can also be seen from the Schwarz approximation (BIC) of the marginal distribution. The Schwarz approximation (see [20, 21]) says that the marginal is asymptotically (up to a constant) $\log(m_\ell(\boldsymbol{x})) \approx \log\left(f_\ell(\boldsymbol{x}|\widehat{\boldsymbol{\theta}}_\ell)\right) - \frac{p}{2}\log(n)$ where p is the dimension of $\boldsymbol{\theta}$. This approximation establishes the consistency of Bayes' Factors for null and nested hypothesis testing through the asymptotic distribution of the log likelihood ratio.

1.2.2 A Paradox from Improper Priors

Though the above discussion suggests that the choice of prior can be asymptotically negligible, it can have a large influence in finite samples and even provide evidence strikingly contradictory to a frequentist analysis. This was first noted by Jeffreys and later termed the Lindley paradox after a landmark paper by Lindley in 1957 [22]. In essence, when testing a point null hypothesis, the evidence for the null can be made arbitrarily large through manipulation of the prior. Lindley showed that while a test of statistical significance can reject the null hypothesis at the α level, the posterior probability of the null hypothesis can be made to be larger than $1-\alpha$. While not too surprising–a poor prior ought to give poor evidence until there is a massive quantity of data–it suggests that care must be taken when specifying priors and using Bayes' Factors. For an example of the paradox in action, consider the Bayes' Factor for 1.1.1.

Example 1.2.1 (Exponential Null Hypothesis Test)**.** *Suppose that $x_1, \ldots, x_n$ are iid $\mathcal{E}(\theta)$ and we have models $M_0: \ \theta = \theta_0$ and $M_1: \ \theta \sim Gamma(\alpha, 1)$ for some $\alpha > 0$. Further assume that the true value of θ is θ^*. The corresponding marginal densities evaluated at $\boldsymbol{x}$ are*

$$m_0(\boldsymbol{x}) = \theta_0^n \exp\left(-n\overline{x}\theta_0\right) \approx \theta_0^n \exp\left(-n\frac{\theta_0}{\theta^*}\right)$$
$$m_1(\boldsymbol{x}) = \frac{\Gamma(n+1)\alpha}{(n\overline{x}+\alpha)^{n+1}} \approx \frac{\sqrt{2\pi}}{\sqrt{n}}(\theta^*)^{n+1}\alpha \exp\left(-n-\alpha\theta^*\right).$$

For fixed α, this provides an asymptotically consistent choice of the true model and $n \to \infty$. However, when n is fixed, allowing $\alpha \to 0$ provides $BF_{01} \to \infty$ regardless of

the value of θ^**. In this example, the posteriors for* M_1 *are consistent for any* α*, even* $\alpha \to 0$ *(which corresponds to an improper uniform prior distribution when computed on compact subsets of* $(0, \infty)$*).*

In 1.2.1 the prior density exhibits mass loss as $\alpha \to 0$. Though the support of each prior is $(0, \infty)$, letting $\alpha \to 0$ produces priors which converge to $\delta_0(\theta)$. Taking this limit creates virtually no change in the posterior distribution of θ for a moderate sample size, but the mass loss is exhibited in the marginal distribution of $\boldsymbol{x}$. If θ^* were indeed 0, then the only data set with any support would have each $x_i = 0$. This type of example can be made with virtually any prior that has open support. Simply let the prior tend to a distribution on the boundary of the set and the null hypothesis is chosen (except in discrete sampling cases where the MLE could lie on the boundary of the set). This not only provides a criticism of proper priors, but also strongly suggests that improper priors cannot be used for comparing model. First, unless one can propose a very good prior, the Bayes' Factor may will provide weights of evidence that are heavily influenced by the prior, regardless of the effect of the prior on the posterior. In fact, priors that exhibit little effect on posterior distributions for small samples exhibit large effect on marginals. For improper priors, in turn, even if the posterior distribution is proper and has "nice" properties, the prior and marginal distributions are only defined up to an arbitrary constant.

1.2.3 Overcoming Issues with Improper Priors

To overcome issues with marginal densities as prior specifications become more vague (or when priors are chosen based on convenience and not elicitation), several alternative model selection criteria can be employed. These include information criteria–like the BIC, which ignores contributions from the prior–and minimal sample methods, which use some notion of training samples. Some of these criteria emerged from an attempt to robustify the marginal distribution and maintain it as the soul arbiter of model quality. Others merely propose a minimization of a particular loss function. While none can truly replace the Bayes' Factor computed from well elicited priors, any criterion that provides the same asymptotic consistency of the Bayes' Factor while reducing prior dependence merits study.

The first class of loss functions seeks to mimic the behavior of a different information criterion, the AIC of Akaike [23, 24]. The AIC seeks to correct the bias in the log likelihood that comes from using the MLE instead of the true value of the parameter. Similar to evaluating the log likelihood at the MLE, Aitkin [25] suggested looking at $\mathbb{E}_{\boldsymbol{\theta}|\boldsymbol{x}}\left[f(\boldsymbol{x}|\boldsymbol{\theta})\right]$ as a measure of model adequacy. This measure allows the use of vague priors, but reflects the same bias as using the plug-in estimator of the MLE. The bias is precisely the dimension of the parameter space, and so–like in the AIC–one can use this as a penalty for the criterion [26] . Similar criteria appear as a term for fit

(often based on deviance) and a term for complexity (as in the BIC, but without the $\log(n)$). The DIC [27] uses the deviance $D(\boldsymbol{\theta}) = -2\log(f(\boldsymbol{x}|\boldsymbol{\theta}))$ to build

$$DIC = \overline{D} + p_D = 2\overline{D} - D(\overline{\boldsymbol{\theta}})$$

where $\overline{D}$ is the posterior expectation of the deviance, $\overline{\boldsymbol{\theta}}$ is the posterior mean, and $p_D = \overline{D} - D(\overline{\boldsymbol{\theta}})$ is an estimation of model complexity. Similarly, Gelfand and Ghosh [28] developed a more generic framework for defining criteria that seeks to minimize a particular predictive loss for a given model. They also obtain a deviance based fit term and a penalty term that is a measure of model complexity.

The second class of methods tries to mimic the marginal distribution. The most basic method takes a Bayesian cross validation approach [29] and looks at a predictive density $\frac{m(\boldsymbol{x})}{m(x_{i_1},\ldots,x_{i_k})}$ where k is the size of a minimal sample (the smallest sample one needs for an improper prior to yield a proper posterior) and all observations are exchangeable. After computing a ratio of these measures, Berger and Pericchi suggest robustifying this measure by taking averages, medians, and geometric means over all possible minimal samples, producing the intrinsic Bayes' Factors (IBF). In addition to these data dependent measures, Berger and Pericchi [30] developed priors based on an asymptotic consideration of this method, termed intrinsic priors. These priors have generated considerable interest among researchers and have been used with great success in a number of problems [31–34] . O'Hagan, in contrast, suggests creating a

measure of model adequacy by dividing out by a fraction of the information in the marginal [35, 36] , defining the fractional marginal as

$$FBF = \frac{m(\boldsymbol{x})}{\int f(\boldsymbol{x}|\boldsymbol{\theta})^b \pi(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}}$$

where $0 < b < 1$. He discusses many choices of b, but $b = \frac{k}{n}$ provides the most clear analogue to the IBF. Beyond these approaches, [37] developed the more general method of expected posterior priors (EPP), defining a prior for each model as

$$\pi_i^*(\boldsymbol{\theta}_i) = \int p_i(\boldsymbol{\theta}_i|\boldsymbol{z}) m^*(\boldsymbol{z}) \mathrm{d}\boldsymbol{z}$$

where m^* is a fixed distribution, often taken to be the marginal from the simplest model in consideration and $\boldsymbol{z}$ is a minimal sample (a sample with the smallest size that provides a proper posterior). In contrast to using the same m^* for all models, Iwaki [38] used the posterior predictive densities of the models to define a data dependent version of this approach. It is important to note that methods like Iwaki's and O'Hagan's are fully Bayesian, but are using some particular loss functions as opposed to only using marginals. The EPP method, on the other hand, is is based purely on marginals when one takes m^* to be a fixed proper distribution which does not depend on the data.

1.2.4 Model Focus and Selection Criteria

When considering statistical models from a fully Bayesian perspective, the notion of what a parameter *is* becomes somewhat muddled. That is to say that in the

Bayesian perspective all analysis can be treated as some sort of missing data problem. Quantities that one knows are treated as known and all other quantities are given distributions that reflect the lack of knowledge of the investigator. In this sense, once a set of distributional assumptions has been made in a problem then any other set of distributional assumptions that lead to the same states of knowledge can be treated as equivalent. In stark contrast to frequency based statistics, where parameters are immutable aspects of nature, the Bayesian treats all quantities distributionally. As prior distributions can play two distinct roles–modeling and estimation–and we can consider parameters at multiple levels in a hierarchy, which parameter one wishes to focus on when analyzing a particular set of data is subject to the views of the individual investigator. When considering a loss function $L(\boldsymbol{a}|\boldsymbol{Y}) = \mathbb{E}_{\boldsymbol{\theta}|\boldsymbol{Y}}\left[L(\boldsymbol{\theta}, \boldsymbol{a}|\boldsymbol{Y})\right]$, the action $\boldsymbol{a}$ is often intimately connected with the parameter $\boldsymbol{\theta}$, as it is when using loss function for choosing posterior summary statistics where the connection is important to the question at hand. However, in the context of model selection, tying a criterion to a specific $\boldsymbol{\theta}$ can become problematic due to the dual role played by priors. When considering a hierarchical model, the choice of level of hierarchy is called the level of model focus [27]. When using a decision criterion that references a specific parameter, the investigator determines the appropriate level of focus for a given analysis. If one chooses this level based on convenience, then the model comparison tool changes dramatically and can fail to correspond to the appropriate model.

Consider a model where we specify sampling distribution $f(\boldsymbol{y}|\boldsymbol{\theta},\boldsymbol{\phi})$ with prior $\pi(\boldsymbol{\theta}|\boldsymbol{\phi})\pi(\boldsymbol{\phi})$ where the priors are integrable. There are then a myriad of choices for sampling distribution (initially denoted as $f(\boldsymbol{y}|\boldsymbol{\theta},\boldsymbol{\phi})$),

$$f(\boldsymbol{y}|\boldsymbol{\phi}) = \int f(\boldsymbol{y}|\boldsymbol{\theta},\boldsymbol{\phi})\pi(\boldsymbol{\theta}|\boldsymbol{\phi})\mathrm{d}\boldsymbol{\theta}$$
$$f(\boldsymbol{y}|\boldsymbol{\theta}) = \int f(\boldsymbol{y}|\boldsymbol{\theta},\boldsymbol{\phi})\pi(\boldsymbol{\phi}|\boldsymbol{\theta})\mathrm{d}\boldsymbol{\phi}.$$

If we consider using a particular level of model focus in an information criterion, we necessarily assume that higher levels in the hierarchy are used only to help estimate parameters and do not represent any additional modeling of the parameters. If we compute the posterior expected log likelihood, we can easily see differences in the criterion.

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\theta},\boldsymbol{\phi}|\boldsymbol{Y}}\left[\log\left(f(Y|\boldsymbol{\theta},\boldsymbol{\phi})\right]\right. &= \log\left(m(\boldsymbol{Y})\right) + \mathbb{E}_{\boldsymbol{\phi},\boldsymbol{\theta}|\boldsymbol{Y}}\left[\log\left(\frac{p(\boldsymbol{\theta},\boldsymbol{\phi}|\boldsymbol{Y})}{\pi(\boldsymbol{\theta},\boldsymbol{\phi})}\right)\right] \\
&= \log\left(m(\boldsymbol{Y})\right) + \mathbb{E}_{\boldsymbol{\phi}|\boldsymbol{Y}}\left[\mathbb{E}_{\boldsymbol{\theta}|\boldsymbol{\phi},\boldsymbol{Y}}\left[\log\left(\frac{p(\boldsymbol{\theta}|\boldsymbol{\phi},\boldsymbol{Y})}{\pi(\boldsymbol{\theta}|\boldsymbol{\phi})}\right)\right]\right] \\
&\quad + \mathbb{E}_{\boldsymbol{\phi}|\boldsymbol{Y}}\left[\log\left(\frac{p(\boldsymbol{\phi}|\boldsymbol{Y})}{\pi(\boldsymbol{\phi})}\right)\right] \\
&= \mathbb{E}_{\boldsymbol{\phi}|\boldsymbol{Y}}\left[\log\left(f(\boldsymbol{Y}|\boldsymbol{\phi})\right)\right] + \mathbb{E}_{\boldsymbol{\phi}|\boldsymbol{Y}}\left[\mathbb{E}_{\boldsymbol{\theta}|\boldsymbol{\phi},\boldsymbol{Y}}\left[\log\left(\frac{p(\boldsymbol{\theta}|\boldsymbol{\phi},\boldsymbol{Y})}{\pi(\boldsymbol{\theta}|\boldsymbol{\phi})}\right)\right]\right] \\
&= \log\left(m(\boldsymbol{Y})\right) + \mathbb{E}_{\boldsymbol{\theta}|\boldsymbol{Y}}\left[\mathbb{E}_{\boldsymbol{\phi}|\boldsymbol{\theta},\boldsymbol{Y}}\left[\log\left(\frac{p(\boldsymbol{\phi}|\boldsymbol{\theta},\boldsymbol{Y})}{\pi(\boldsymbol{\phi}|\boldsymbol{\theta})}\right)\right]\right] \\
&\quad + \mathbb{E}_{\boldsymbol{\theta}|\boldsymbol{Y}}\left[\log\left(\frac{p(\boldsymbol{\theta}|\boldsymbol{Y})}{\pi(\boldsymbol{\theta})}\right)\right] \\
&= \mathbb{E}_{\boldsymbol{\theta}|\boldsymbol{Y}}\left[\log\left(f(\boldsymbol{Y}|\boldsymbol{\theta})\right)\right] + \mathbb{E}_{\boldsymbol{\theta}|\boldsymbol{Y}}\left[\mathbb{E}_{\boldsymbol{\phi}|\boldsymbol{\theta},\boldsymbol{Y}}\left[\log\left(\frac{p(\boldsymbol{\phi}|\boldsymbol{\theta},\boldsymbol{Y})}{\pi(\boldsymbol{\phi}|\boldsymbol{\theta})}\right)\right]\right]
\end{aligned}$$

As is easily seen, each posterior expected log likelihood appears as the "full" value minus some bias term. This type of analysis–which applies to the DIC–can be extended to provide critiques of the FBF and Gelfand's method.

Invariance to issues of model focus is an additional reason for why the marginal distribution is afforded a unique role in model selection. Marginal distributions, as well as distributions derived from marginal distributions, are the only objects that are invariant to the choice of model focus. In fact, since posterior predictive distributions are defined in terms of ratios of marginal distributions ($pr(\boldsymbol{Z}|\boldsymbol{Y}) = \frac{m(\boldsymbol{Z},\boldsymbol{Y})}{m(\boldsymbol{Y})}$) they are also invariant to choice of model focus. One must exercise caution, however, when defining posterior predictive densities as they are often defined in terms of conditional independence with respect to a particular parameter ($\boldsymbol{Z} \perp\!\!\!\perp \boldsymbol{Y}|\boldsymbol{\theta}$). One then defines the posterior predictive density as

$$pr(\boldsymbol{Z}|\boldsymbol{Y}) = \int f(\boldsymbol{Z}|\boldsymbol{\theta},\boldsymbol{Y})p(\boldsymbol{\theta}|\boldsymbol{Y})\mathrm{d}\boldsymbol{\theta} = \int f(\boldsymbol{Z}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{Y})\mathrm{d}\boldsymbol{\theta}.$$

The particular choice of conditional independence implies the definitions of conditional sampling densities. In particular, we have $f(\boldsymbol{Z}|\boldsymbol{\theta},\boldsymbol{Y}) = f(\boldsymbol{Z}|\boldsymbol{\theta})$ and

$$f(\boldsymbol{Z}|\boldsymbol{\theta},\boldsymbol{\phi},\boldsymbol{Y}) = \frac{p(\boldsymbol{\phi}|\boldsymbol{\theta},\boldsymbol{Z},\boldsymbol{Y})f(\boldsymbol{Z}|\boldsymbol{\theta},\boldsymbol{Y})}{p(\boldsymbol{\phi}|\boldsymbol{\theta},\boldsymbol{Y})} = \frac{p(\boldsymbol{\phi}|\boldsymbol{\theta},\boldsymbol{Z},\boldsymbol{Y})f(\boldsymbol{Z}|\boldsymbol{\theta})}{p(\boldsymbol{\phi}|\boldsymbol{\theta},\boldsymbol{Y})}$$
$$f(\boldsymbol{Z}|\boldsymbol{\phi},\boldsymbol{Y}) = \int f(\boldsymbol{Z}|\boldsymbol{\theta},\boldsymbol{\phi},\boldsymbol{Y})p(\boldsymbol{\theta}|\boldsymbol{\phi},\boldsymbol{Y})\mathrm{d}\boldsymbol{\theta},$$

and it is easy to verify that they all lead to the same posterior predictive density. While the initial choice of conditional independence appears to be tied to a choice of model focus, the choices differ. When choosing a conditional independence re-

lationship, the investigator is envisioning a particular sampling scheme for future observations from the process under consideration. As such, the particular choice of independence relationship is tied to a data-based inferential problem, and not to the whims of the investigator or the convenience of a specific hierarchical level. After the predictive inferential goal is established, the posterior predictive density is fixed. It is in this way that posterior predictive distributions can be viewed as being independent of the choice of model focus. Of the methods discussed, the BF, IBF, EPP, and Iwaki's method are invariant to issues of model focus.

1.3 Information Theory

In this chapter, we discuss some general results in information theory that motivate the selection criteria developed in the dissertation. Historically, information theory began with the work of Claude Shannon. Beyond its applications in signal processing (for which it was introduced), information theory is now used regularly across science and engineering as well as in statistics. For example, the Bernardo rule for deriving a reference prior uses the quantity

$$I = \int \int \log \left(\frac{p(\boldsymbol{\theta}|\boldsymbol{x}_n)}{\pi(\boldsymbol{\theta})} \right) p(\boldsymbol{\theta}|\boldsymbol{x}_n) \mathrm{d}\boldsymbol{\theta} m(\boldsymbol{x}_n) \mathrm{d}\boldsymbol{x}_n$$

which is the expected mutual information between $\boldsymbol{\theta}$ and $\boldsymbol{x}_n$. The mutual information in a channel was first defined in Shannon's groundbreaking paper [39] and minimizing this mutual information provides the most efficient encoding for the channel. Though there is a deep geometry involved in information theory [40–44] , this geometry is

mostly tied to estimation problems and the manifolds developed are manifolds for particular families of distributions (for example, a particular exponential distribution where the manifold is defined in terms of the parameters of the exponential family). Since we are performing statistical estimation in terms of placing a prior distribution over such a manifold, the geometry is most interesting when defining a prior distribution. When considering model selection, we would like to build criteria that view models only through observable data in order to avoid issues of model focus. Thus we restrict out attention to measures of divergence between distributions which can be applied to predictive and marginal densities.

1.3.1 Shannon Entropy and K-L Divergence

Suppose that we have a discrete distribution given by $\boldsymbol{p} = (p_1, \ldots, p_n)$ and define a function $H(\boldsymbol{p})$ by the properties [45]

1. $H(p_1, \ldots, p_n)$ is symmetric in its arguments
2. $H(p, 1-p)$ is continuous for $p \in [0, 1]$
3. $H(.5, .5) = \log(2)$
4. $H(tp_1, (1-t)p_1, p_2, \ldots, p_n) = H(p_1, \ldots, p_n) + p_1 H(t, 1-t)$ for $t \in [0, 1]$

then the unique function H, called the Shannon Entropy, is

$$H(p_1, \ldots, p_n) = \sum p_i \log\left(\frac{1}{p_i}\right)$$

Proof. The proof of this fact is reduced to the following, due to Erdös (see Rényi [45] for a proof of Erdös' theorem). If f is a function defined on $\mathbb{N}$ such that $f(nm) = f(n) + f(m)$ and $\lim_{n\to\infty} f(n+1) - f(n) = 0$ then $f(n) = c\log(n)$. In order to use this to prove the form of H, we can define an appropriate f and use continuity. Define

$$f(n) = H\left(\frac{1}{n}, \ldots, \frac{1}{n}\right).$$

It is easy to see that the additivity condition extends to $t_1, \ldots, t_n$ in the $n-$simplex using induction. In general, the additivity condition can be further extended to break down the measure of information into one over all measure of information plus a conditional expectation. Define R latent groupings as $G_r = \{k_r+1, \ldots, k_{r+1}\}$ where $k_1 = 1$ and $k_{R+1} = n$, the probability of being in group G_r as $\omega_r = \sum_{i=k_r+1}^{k_{r+1}} p_i$, and $\boldsymbol{p}_r$ to be the vector formed by $\frac{p_i}{\omega_r}$ for $i \in G_r$. The function H satisfies

$$H(\boldsymbol{p}) = H(\boldsymbol{\omega}) + \sum_{r=1}^{R} H(\boldsymbol{p}_r)\omega_r.$$

In fact, this equation is equivalent to the simple additivity condition and so the Shannon entropy will be the unique information function with this additivity property. To see how this extended additivity property implies leads to the Shannon entropy, consider $\boldsymbol{p}$ to be a vector of length mn with each element being $\frac{1}{mn}$. Define the groupings by $k_r = (r-1)n$ for $r = 1, \ldots, m$, then each ω_r is $\frac{1}{m}$ and each $\boldsymbol{p}_r$ is a vector of length n with each element being $\frac{1}{n}$. Thus, the additivity condition gives us $f(mn) = f(m) + f(n)$. Now, we have to show that $\lim_{n\to\infty} f(n+1) - f(n) = 0$ and

we will be able to apply Erdös' theorem. Using the additivity condition, we can see that

$$f(n+1) = H\left(\frac{n}{n+1}, \frac{1}{n+1}\right) + \frac{n}{n+1} f(n).$$

Because $H(0,1) = 0$ and the continuity assumption, $H\left(\frac{n}{n+1}, \frac{1}{n+1}\right) \to 0$. It follows that $f(n+1) - f(n) \to 0$ and $f(n) = c \log((\,) n)$. Since $H(.5, .5) = f(2) = \log(2)$, we can conclude that $c = 1$ □

The interpretation of the Shannon entropy is fairly straightforward. The amount of information from observing outcome i is $\log\left(\frac{1}{p_i}\right)$ and so the Shannon entropy is the expected information of the probability vector. Because $0 \leq p_i \leq 1$ for all i, it is easy to see that the Shannon entropy is positive, with the value being 0 if and only if the probability vector contains a single 1. It is an easy calculus problem to show that $\sum p_i \log\left(\frac{1}{q_i}\right)$ is minimized when $q_i = p_i$ for all i. And so, given two probability vectors, we can define the average information difference of $\boldsymbol{p}$ from $\boldsymbol{q}$ to be

$$D_{KL}(p||q) = \sum_{i=1}^{n} p_i \log\left(\frac{p_i}{q_i}\right).$$

This is the divergence introduced to Kullback and Leibler [46] and represents the information gain when $\boldsymbol{q}$ is proposed as the probability distribution and $\boldsymbol{p}$ is the actual probability distribution. In this way, the Shannon entropy can be viewed as the information gain of $\boldsymbol{p}$ from the uniform distribution.

1.3.2 Rényi Divergences

In order to generalize the idea of information gain when one distribution is replaced by another, Rényi [45] proposed replacing the average of the logarithms with a more generic average

$$D_g(\boldsymbol{p}||\boldsymbol{q}) = g^{-1}\left[\sum_i p_i g\left(\log\left(\frac{p_i}{q_i}\right)\right)\right]$$

where g is a strictly increasing function If we define $\boldsymbol{p}_1 \star \boldsymbol{p}_2$ to be the distribution induced by independence, then Rényi also requires additivity:

$$D_g(\boldsymbol{p}_1 \star \boldsymbol{p}_2||\boldsymbol{q}_1 \star \boldsymbol{q}_2) = D_g(\boldsymbol{p}_1||\boldsymbol{q}) + D_g(\boldsymbol{p}_2||\boldsymbol{q}_2). \tag{1.7}$$

This requirement is actually quite strict and g has to satisfy $g(x+y) = g(x) + a(x)g(y) = g(y) + a(y)g(x)$. Thus for all x and y, which induces the equality $a(x) = 1 + kg(x)$ for all x. In the case where $k = 0$, one obtains $g(x+y) = g(x) + g(y)$ and so g is a linear function and the K-L divergence is obtained. However, for $k \neq 0$, we have

$$\frac{a(x+y)-1}{k} = g(x) + a(x)g(y) = \frac{a(x)-1}{k} + a(x)\frac{a(y)-1}{k},$$

which implies that $a(x+y) = a(x)a(y)$. The monotonicity assumption tells us that $a(x) = c\exp((\alpha-1)x)$ and so $g(x) = \frac{1}{k}(c\exp((\alpha-1)x) - 1)$. Substituting this back into (1.7), we get

$$D_\alpha(\boldsymbol{p}||\boldsymbol{q}) = \frac{1}{\alpha-1}\log\left(\left[\sum p_i\left(\frac{p_i}{q_i}\right)^{\alpha-1}\right]\right).$$

For a continuous probability space with measures P and Q, define $h = \frac{\mathrm{d}P}{\mathrm{d}Q}$ when $Q \ll P$ and define

$$D_\alpha(P||Q) = \frac{1}{\alpha - 1} \log \left(\int h^{\alpha-1} dP \right)$$
$$D_\alpha(p||q) = \frac{1}{\alpha - 1} \log \left(\int \left(\frac{p(x)}{q(x)} \right)^{\alpha-1} p(x) \mathrm{d}x \right) \tag{1.8}$$

where p and q and the densities of P and Q when both are absolutely continuous with respect to the measure given by $\mathrm{d}x$. It is easy to see that these measures of divergence are continuous in $\alpha > 0$.

In order to state some simple properties of the Rényi divergences, we need access to Jensen's inequality. If ϕ is a convex function on $\mathbb{R}$ and g is a μ integrable function for a probability measure μ, then $\phi\left(\int g \mathrm{d}\mu\right) \leq \int \phi(g) \mathrm{d}\mu$.

If we consider the function $\phi(x) = x^{\frac{\alpha-1}{\alpha'-1}}$ for $\alpha > \alpha'$ (where neither is 0), then ϕ is convex. Thus $D_{\alpha'}(P||Q) \leq D_\alpha(P||Q)$ whenever $\alpha' \leq \alpha$. When $0 < \alpha < 1$, we can use the fact that

$$D_\alpha(p||q) = \frac{1}{\alpha - 1} \log \left(\int \left(\frac{p(x)}{q(x)} \right)^{\alpha-1} p(x) \mathrm{d}x \right) = -\frac{1}{1 - \alpha} \log \left(\int \left(\frac{p(x)}{q(x)} \right)^{\alpha} q(x) \mathrm{d}x \right)$$

and that $\phi(x) = x^\alpha$ is concave to conclude the fact that
$\int \left(\frac{p(x)}{q(x)}\right)^\alpha q(x)\mathrm{d}x < \left(\int q(x)\mathrm{d}x\right)^\alpha = 1$ and so $D_\alpha(p||q) > 0$ for all α.

A few values of α provide straightforward interpretation

- $D_\infty(p||q) = \log\left(\mathrm{ess.sup}_x \frac{p(x)}{q(x)}\right)$: often infinite
- $D_2(p||q) = \log\left(\int \frac{p(x)^2}{q(x)}\mathrm{d}x\right)$: the log expected ratio of the densities (related to χ^2)

- $D_1(p||q) = \int \log\left(\frac{p(x)}{q(x)}\right) p(x)\mathrm{d}x$: the K-L divergence
- $D_{\frac{1}{2}}(p||q) = -2\log\left(\int \sqrt{p(x)q(x)}\mathrm{d}x\right)$: twice the log affinity (related to Hellinger)
- $D_0(p||q) = -\log\left(\int_{\mathrm{supp}(p(x))} q(x)\mathrm{d}x\right)$: log measure of common support

Intuitively, the α divergence depends on lower probability outcomes (under p) as α becomes smaller. To see the connection between D_2 and the Pearson χ^2 distance, notice that the Pearson distance is

$$\int \left(\frac{p(x)}{q(x)} - 1\right)^2 q(x)\mathrm{d}x = \exp\left(D_2(p||q)\right) - 1.$$

To see the connection between $D_{\frac{1}{2}}$ and the Hellinger distance, notice that the Hellinger distance is

$$\int \left(\sqrt{p(x)} - \sqrt{q(x)}\right)^2 \mathrm{d}x = 2\left(1 - \exp\left(D_{\frac{1}{2}}(p||q)\right)\right)$$

In fact, we have already seen a number of Rényi divergences. The most obvious place where they are being used is when computing the posterior expected log-likelihood for the DIC,

$$\begin{aligned}\int \log\left(f(\boldsymbol{x}|\boldsymbol{\theta})\, p(\boldsymbol{\theta}|\boldsymbol{x})\mathrm{d}\boldsymbol{\theta}\right. &= \log\left(m(\boldsymbol{x})\right) + \int \log\left(\frac{p(\boldsymbol{\theta}|\boldsymbol{x})}{\pi(\boldsymbol{\theta})}\right) p(\boldsymbol{\theta}|\boldsymbol{x})\mathrm{d}\boldsymbol{\theta} \\ &= \log\left(m(\boldsymbol{x})\right) + D_1(p(\cdot|\boldsymbol{x})||\pi) \\ &= \log\left(m(x)\exp\left(D_1(p(\cdot|\boldsymbol{x})||\pi)\right)\right).\end{aligned}$$

They also appear in Aitkin's posterior Bayes' Factor,

$$\int f(\boldsymbol{x}|\boldsymbol{\theta}p(\boldsymbol{\theta}|\boldsymbol{x})\mathrm{d}\boldsymbol{\theta} = m(\boldsymbol{x})\int \frac{p(\boldsymbol{\theta}|\boldsymbol{x})^2}{\pi(\boldsymbol{\theta})}\mathrm{d}\boldsymbol{\theta} = m(\boldsymbol{x})\exp\left(D_2(p(\cdot|\boldsymbol{x})||\pi)\right).$$

Most notably, they appear in O'Hagan's FBF because the fractional marginal is

$$\begin{aligned}\frac{m(\boldsymbol{x})}{\int f(\boldsymbol{x}|\boldsymbol{\theta})^b \pi(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}} &= \frac{m(\boldsymbol{x})}{\int m(\boldsymbol{x}) \left(\frac{f(\boldsymbol{\theta}|\boldsymbol{x})}{\pi(\boldsymbol{\theta})}\right)^b \pi(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}} \\ &= \frac{m(\boldsymbol{x})^{1-b}}{\int \left(\frac{f(\boldsymbol{\theta}|\boldsymbol{x})}{\pi(\boldsymbol{\theta})}\right)^{b-1} f(\boldsymbol{\theta}|\boldsymbol{x})\mathrm{d}\boldsymbol{\theta}} \\ &= \left(m(\boldsymbol{x}) \exp\left(D_b(p(\cdot|\boldsymbol{x})||\pi)\right)\right)^{1-b}\end{aligned}$$

All three of these measures of model adequacy appear as the marginal modified by a Rényi entropy of the posterior distribution from the prior distribution for the parameters. It is now apparent why these measures behave differently. The fit term from the DIC and the posterior Bayes' Factor are using Rényi entropies with a fixed α. As we collect more and more data, if we have a sequence of consistent posteriors, then eventually the Rényi divergences become infinite. In contrast, the FBF maintains a check on the growth of the divergence by forcing $\alpha = \alpha(n) \to 0$. In this way, the FBF produces behavior similar to the Bayes' Factor, whereas the other two quantities allow for model comparison to select the alternative whenever the null is true, even as the amount of data becomes infinite.

In the following chapter, we wish to build criteria using information integrals in the family the Rényi introduced. One family will be purely based on K-L divergences and the other will be built on the entire family of Rényi divergences. One important aspect of the latter will be its asymptotic correspondence to Bayes' Factors. For entire families for any value of α, behavior like that of a Bayes' Factor can be achieved when one has access to minimal samples and predictive densities. This is in stark contrast

to the family of posterior divergences exhibited above, as the only ones that can possibly exhibit consistent model selection are the FBFs, and consistency is not even guaranteed for them.

2. New Alternatives for Model Selection

In this chapter, we will introduce some new families of information criteria to be used for model comparison and selection. The first is a criterion based purely on predictive densities and the Kullback-Leibler Divergence. This measure will include a fit term based on the posterior Bayes' Factor (although any measure that behaves in a similar fashion would suffice) and a term which penalizes for complexity by looking at the relative K-L Divergences of the two models. This can be extended in a number of ways, which are discussed but not explored fully. The second family is based upon modifying the marginal distribution by an appropriate Rényi divergence. This is a large and flexible family of criteria that provide asymptotic selection properties of the Bayes' factor while allowing the use of diffuse and improper priors. It is clearly seen that this family provides analogues to the intrinsic and fractional Bayes' factor by placing Iwaki's method in a larger family of criteria.

2.1 Posterior Predictive Information Criterion

The Signed Divergence

Consider two distributions f and g for data $\boldsymbol{x}$. We want to find an intrinsic measure which provides a natural definition of of the relative quality of the two models in

terms of information for discrimination $I(f,g)$. A natural choice is one of the K-L divergences $D(f||g)$ or $D(g||f)$ or the Jensen-Shannon divergence $D(f||g)+D(g||f)$. The only problem is that each of these measures is positive and so gives us no notion of which model provides more discrimination. In order to do this, it is reasonable to look at the difference in K-L divergences, defining a new signed divergence here by

$$I(f,g) = D(f \parallel g) - D(g \parallel f),$$

which is clearly anti-symmetric in its variables. In contrast to the Jensen-Shannon divergence which would take the sum of the K-L Divergences, we focus on taking the difference in order to determine which model provides more information for discrimination. Though a somewhat vague notion itself, this measure of divergence should favor, in terms of sign, the distribution which is more diffuse. This is because the value of a K-L divergence is strongly influenced by sets of small measure. In fact, the signed divergence can be viewed as a difference in two other K-L divergences:

$$\begin{aligned}
I(f,g) &= D(f \parallel g) - D(g \parallel f) \\
&= \int \log\left(\frac{f}{g}\right) f - \int \log\left(\frac{g}{f}\right) g \\
&= \int \log\left(\frac{f}{g}\right) f + \int \log\left(\frac{f}{g}\right) g \\
&= 2\int \log\left(\frac{f}{g}\right) \frac{f+g}{2} \\
&= 2\left[\int \log\left(\frac{2f}{f+g}\right) \frac{f+g}{2} + \int \log\left(\frac{f+g}{2g}\right) \frac{f+g}{2}\right] \\
&= 2\left[D(h \parallel g) - D(h \parallel f)\right]
\end{aligned}$$

where h is the probability density $\frac{f+g}{2}$. The sign of this divergence favoring a more diffuse distribution can now be viewed as saying that the distribution which is more diffuse is the distribution which is closer to the average of the two distributions. This notion of diffuseness can be solidified by looking at two classical examples.

Proposition 2.1.1. *Suppose the F is a Bernoulli distribution with parameter q and that G is a Bernoulli distribution with parameter p. Further suppose that* $\max\{p, 1-p, q, 1-q\} = \max\{q, 1-q\}$. *Then $I(F,G) \leq 0$ and $I(F,G) = 0$ if and only if $p = q$ or $p = 1-q$.*

Proof. Without loss of generality, we can assume that $q = \max\{q, 1-q\}$ and thus assume $q \geq \frac{1}{2}$. Define

$$I_q(p) = I(F,G) = (p+q)\log\left(\frac{q}{p}\right) + (2-p-q)\log\left(\frac{1-q}{1-p}\right)$$

For fixed q, it is easy to see that $I_q \in C^\infty(0,1)$ and $I_q(1-q) = 0 = I_q(q)$. Thus, $I(F,G) = 0$ whenever $q = \frac{1}{2}$. Suppose the $q > \frac{1}{2}$. The first two derivatives of I_q are:

$$I_q'(p) = \log\left(\frac{q}{p}\right) - \log\left(\frac{1-q}{1-p}\right) - \frac{q}{p} + \frac{1-q}{1-p}$$

$$I_q''(p) = -\frac{1}{p} - \frac{1}{1-p} + \frac{q}{p^2} + \frac{1-q}{(1-p)^2}$$

Algebraic manipulation shows that

$$I_q''(p) = \frac{(1-2p)(q-p)}{p^2(1-p)^2}$$

It is easy to see that $I_q'(q) = 0$ and that $I_q''(p) = 0$ has only two solutions, one at $p = q$ and one at $p = \frac{1}{2}$. If $I_q(p)$ were to have more roots than $p = q$ and $p = 1-q$,

then I_q would have to have more inflection points. And so the sign of $I_q'(1-q)$ will tell us the sign of $I_q(p)$ for $p \in (1-q, q)$. Define $T(q) = I_q'(q)$. It is easy to see that $T(\frac{1}{2}) = 0$. Also,

$$T'(q) = (2q-1)\left(-\frac{1}{(1-q)^2} - \frac{1}{q^2}\right)$$

which is negative for all $q > \frac{1}{2}$. Thus, $T(q) < 0$ for all $q > \frac{1}{2}$ and we can conclude that $I_q(p) < 0$ for all $1 - q < p < q$. □

Remark 2.1.1. *We can easily compute the signed divergence for two univariate normal distributions. Suppose that f is a normal density with mean μ and precision s and the g is a normal density with mean θ and precision t.*

$$\begin{aligned}
D(f \parallel g) &= \mathbb{E}_f\left[\frac{1}{2}\log\left(\frac{s}{t}\right) - \frac{s}{2}(y-\mu)^2 + \frac{t}{2}(y-\theta)^2\right] \\
&= \frac{1}{2}\log\left(\frac{s}{t}\right) - \frac{s}{2}\mathbb{E}_f\left[(y-\mu)^2\right] + \frac{t}{2}\mathbb{E}_f\left[(y-\theta)^2\right] \\
&= \frac{1}{2}\log\left(\frac{s}{t}\right) - \frac{1}{2} + \frac{t}{2}\mathbb{E}_f\left[(y-\theta)^2\right] \\
&= \frac{1}{2}\log\left(\frac{s}{t}\right) - \frac{1}{2} + \frac{t}{2}\mathbb{E}_f\left[(y-\mu+\mu-\theta)^2\right] \\
&= \frac{1}{2}\log\left(\frac{s}{t}\right) - \frac{1}{2} + \frac{t}{2}\mathbb{E}_f\left[(y-\mu)^2\right] + \frac{t}{2}(\mu-\theta)^2 \\
&= \frac{1}{2}\log\left(\frac{s}{t}\right) - \frac{1}{2} + \frac{t}{2s} + \frac{t}{2}(\mu-\theta)^2
\end{aligned}$$

Symmetry dictates that

$$D(g \parallel f) = \frac{1}{2}\log\left(\frac{t}{s}\right) - \frac{1}{2} + \frac{s}{2t} + \frac{s}{2}(\theta-\mu)^2$$

Thus the signed divergence is

$$I(f,g) = \log\left(\frac{s}{t}\right) + \frac{1}{2}\left(\frac{t}{s} - \frac{s}{t}\right) + \frac{1}{2}(\mu-\theta)^2(t-s)$$

Proposition 2.1.2. *Suppose that f is a normal density with mean μ and precision s and that g is a normal density with mean θ and precision t. Further suppose that $s \geq t$. Then $I(f,g) \leq 0$ and $I(f,g) = 0$ if and only if $s = t$.*

Proof. Fix $s > 0$ and define $x = \frac{t}{s}$. Define

$$I_s(x) = I(f,g) = \log(x) + \frac{1}{2}\left(x - \frac{1}{x}\right) + \frac{s}{2}(\mu - \theta)^2(x-1)$$

For fixed s, $I_s(x) \in C^\infty(0,\infty)$. Trivially $I_s(1) = 0$ so we have shown that $I(f,g) = 0$ whenever $s = t$. Also,

$$I_s'(x) = \frac{1}{x} + \frac{1}{2}\left(1 + \frac{1}{x^2}\right) + \frac{s}{2}(\mu - \theta)^2 > 0$$

If $s > t$ then $x < 1$ and $I_s'(x) > 0$ together with $I_s(1) = 0$ imply that $I(f,g) < 0$. □

Remark 2.1.2. *The measure I can be easily computed for multivariate normal distributions. Let f be a normal density in d dimensions with mean $\boldsymbol{\mu}$ and precision matrix $\boldsymbol{S}$. Let g be a normal density in d dimensions with mean $\boldsymbol{\theta}$ and precision matrix $\boldsymbol{T}$. Some algebra shows that*

$$I(f,g) = \log\left(\frac{|\boldsymbol{S}|}{|\boldsymbol{T}|}\right) + \frac{1}{2}tr\left(\boldsymbol{S}^{-1}\boldsymbol{T} - \boldsymbol{T}^{-1}\boldsymbol{S}\right) + \frac{1}{2}(\boldsymbol{\mu} - \boldsymbol{\theta})^T(\boldsymbol{T} - \boldsymbol{S})(\boldsymbol{\mu} - \boldsymbol{\theta}). \quad (2.1)$$

This is analogous to the measure for univariate normal distributions. In fact, whenever $\boldsymbol{T} = t\boldsymbol{I}$, $\boldsymbol{S} = s\boldsymbol{I}$, $\boldsymbol{\mu} = \mu_0\mathbf{1}$, and $\boldsymbol{\theta} = \theta_0\mathbf{1}$, where $\boldsymbol{I}$ is the $d \times d$ identity matrix and $\mathbf{1}$ is a length d vector of ones, the measure collapses to d times what one would get with the corresponding univariate measure.

These examples help solidify what we mean by saying one distribution is more diffuse than another. In the Bernoulli example, the signed divergence follows the distribution which has parameter furthest from $\frac{1}{2}$, which corresponds to the distribution with larger standard deviation. In the Gaussian example, the signed divergence follows the distribution with larger standard deviation. We will use this measure when f and g are the posterior predictive densities of two different models. This measure should then favor (in terms of sign) the model which is more complex because integrating out the parameters will make that particular predictive density less narrow. Essentially, one of the nicest aspects of Bayesian statistics will provide leverage over the models: integration to marginalize out a variable creates an automatic Occam's razor.

2.2 The PPIC

Suppose that we have two models M_i, $i = 1, 2$. That is, suppose that we have sampling distributions $F_i(\boldsymbol{y}|\boldsymbol{\theta}_i)$ and priors $\Pi_i(\boldsymbol{\theta}_i)$ such that the Bayesian predictive distributions $F_i(\boldsymbol{y}|\boldsymbol{Y})$ (from collecting data $\boldsymbol{Y}$) are mutually absolutely continuous. We can find a measure μ with respect to which the $F_i(\boldsymbol{y}|\boldsymbol{Y})$ are both absolutely continuous. Define the predictive densities (with respect to μ) as $f_i(\boldsymbol{y}|\boldsymbol{Y})$ where $\boldsymbol{y}$ is data from a fully repeated experiment. When predictive densities have common support we can define the following information theoretic quantity:

Definition 2.2.1. *Define the quantity*

$$W_\alpha(M_1, M_2) = \log\left(\frac{f_1(\boldsymbol{Y}|\boldsymbol{Y})}{f_2(\boldsymbol{Y}|\boldsymbol{Y})}\right) - \alpha I(f_1, f_2) \tag{2.2}$$

What W_α represents is a term for fit (the log-ratio of the predictive distributions) minus a term that takes into account the difference in average predictive discrimination of the two models. Since the second term favors models with more spread and more complex models have more spread due to integrating out parameters, we can see that subtracting the signed divergence provides a check on overfitting in complex models. The next proposition shows a simple consistency property of the measure W_α.

Proposition 2.2.1. *Suppose* $M_i : y_j \overset{iid}{\sim} f_i(x|\boldsymbol{\theta}_i)$ *for* $j = 1, \ldots, n$ *and* $i = 1, 2$ *where the* $\boldsymbol{\theta}_i$ *are fixed. Then* $W_\alpha(M_1, M_2)$ *is asymptotically consistent for all* $\alpha \in (0, 1)$.

Proof. Suppose that M_1 is the true model, then

$$\begin{aligned}
\log\left(\frac{f_1(y_1, \ldots y_n|\boldsymbol{\theta}_1)}{f_2(y_1, \ldots y_n|\boldsymbol{\theta}_2)}\right) &= \log\left(\prod_{j=1}^{n} \frac{f_1(y_j|\boldsymbol{\theta}_1)}{f_2(y_j|\boldsymbol{\theta}_2)}\right) \\
&= \sum_{j=1}^{n} \log\left(\frac{f_1(y_j|\boldsymbol{\theta}_1)}{f_2(y_j|\boldsymbol{\theta}_2)}\right) \\
&\approx n D_1(f_1 \parallel f_2)
\end{aligned}$$

where D_1 is the Kullback-Leibler Divergence for an observation set of cardinality 1. It is easy to see that

$$I(f_1, f_2) = n\left[D_1(f_1 \parallel f_2) - D_1(f_2 \parallel f_1)\right]$$

Thus, we can conclude that

$$W_\alpha(M_1, M_2) \approx n\left(\alpha D_1(f_2 \parallel f_1) + (1-\alpha) D_1(f_1 \parallel f_2)\right) > 0$$

and we would select M_1 asymptotically. Asymmetry provides the result that M_2 being the correct model gives $W < 0$ and M_2 would be chosen asymptotically. $\square$

Remark 2.2.1. *There is an interesting observation about the asymptotic form of $W_{\frac{1}{2}}$ from this proposition. When g has actually generated the independent data, then $W_{\frac{1}{2}}(M_1, M_2) \approx n \int \left(g - \frac{f_1+f_2}{2}\right) \log\left(\frac{f_1}{f_2}\right)$. The criteria selects precisely the model which is has less divergence separating it from the data generating process than from the mixture model of the two models with equal weightings. This may suggest that a choice of $\alpha = \frac{1}{2}$ is appropriate for the test, but we will see that may not penalize complex models enough to get consistency, although it will still provide a check on complexity, much like that in the AIC or DIC.*

2.2.1 Examples

Proposition 2.2.2. *The W_α test can be made to have arbitrarily small Type I error rate for a point null test of the mean of independent normally distributed data when the precision is known and $\alpha \in \left(\frac{1}{2}, 2\right)$*

Proof. Suppose that $M_1 : Y_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$ and that $M_2 : Y_i \overset{\text{iid}}{\sim} \mathcal{N}(\mu, 1)$ with prior $\pi(\mu) = c$. Then the posterior for μ is $\mu \sim \mathcal{N}(\overline{\boldsymbol{Y}}, 1/n)$ under M_2. The predictive

density under M_2 for a replicated experiment is determined by working through the quadratic form:

$$\begin{aligned} Q &= (\boldsymbol{y} - \mu\mathbf{1})^T(\boldsymbol{y} - \mu\mathbf{1}) + n(\mu - \overline{\boldsymbol{Y}})^T(\mu - \overline{\boldsymbol{Y}}) \\ &= \mu\left(1+n\right)\mu - 2\mu\left(\mathbf{1}^T\boldsymbol{y} + n\overline{\boldsymbol{Y}}\right) + f(\boldsymbol{Y}) \\ &= Q_\mu + \left(y - \frac{1}{n}\boldsymbol{J}\boldsymbol{Y}\right)^T\left(\boldsymbol{I} - \frac{1}{2n}\boldsymbol{J}\right)\left(\boldsymbol{y} - \frac{1}{n}J\boldsymbol{Y}\right) + f(\boldsymbol{Y}) \end{aligned}$$

where I is the $n \times n$ identity matrix, J is an $n \times n$ matrix of $1's$, $\mathbf{1}$ is a vector of $1's$ of length n, Q_μ is a quadratic form in μ, $f(\boldsymbol{Y})$ is some function of $\boldsymbol{Y}$, and n is the size of the data set. Thus the predictive density for $\boldsymbol{y}$ given that we have observed data $\boldsymbol{Y}$ under model M_2 is multivariate normal with mean $\overline{\boldsymbol{Y}}\mathbf{1}$ and precision matrix $\boldsymbol{I} - \frac{1}{2n}\boldsymbol{J}$.

Now we need to compute the predictive densities at the data itself

$$f_1(\boldsymbol{Y}|\boldsymbol{Y}) = \left(\frac{1}{\sqrt{2\pi}}\right)^n \exp\left(-\frac{1}{2}\boldsymbol{Y}^T\boldsymbol{Y}\right)$$

$$\begin{aligned} f_2(\boldsymbol{Y}|\boldsymbol{Y}) &= \frac{\left(\det\left[\boldsymbol{I} - \frac{1}{2n}\boldsymbol{J}\right]\right)}{(\sqrt{2\pi})^n}\exp\left(-\frac{1}{2}\left(\boldsymbol{Y} - \overline{\boldsymbol{Y}}\mathbf{1}\right)^T\left(\boldsymbol{I} - \frac{1}{2n}\boldsymbol{J}\right)\left(\boldsymbol{Y} - \overline{\boldsymbol{Y}}\mathbf{1}\right)\right) \\ &= \frac{1}{2(\sqrt{2\pi})^n}\exp\left(-\frac{1}{2}\left(\boldsymbol{Y}^T\boldsymbol{Y} - n\overline{\boldsymbol{Y}}^2\right)\right) \end{aligned}$$

This provides a measure for discrimination of

$$W_\alpha(M_1, M_2) = \frac{1}{2}\log\left(2\right) - \frac{n}{2}\overline{\boldsymbol{Y}}^2 - \alpha\log\left(2\right) + \frac{3\alpha}{4} + \frac{n\alpha}{4}\overline{\boldsymbol{Y}}^2$$

Under the assumption that M_1 is true, we know that $n\overline{\boldsymbol{Y}}^2$ has a chi-squared distribution with 1 degree of freedom, we can see that the probability that $W(M_1, M_2) < 0$

is the probability that $n\overline{\boldsymbol{Y}}^2 < \frac{3\alpha - 2\log(2)(2\alpha-1)}{2-\alpha}$. This probability is greater than .5 whenever $\alpha > -0.7$. Additionally, it is easy to see that this probability grows to 1 as α approaches 2. Choice of α in this range provides control of Type I error.

When M_2 is the correct model, we know that $n\overline{\boldsymbol{Y}}^2$ has a non-central chi-squared distribution with 1 degree of freedom and non-centrality parameter $n(\mu^*)^2$, where μ^* is the actual value of μ generating the data. We can easily see that the probability that $W_\alpha < 0$ increases to 1 as n increases for any $\alpha < 2$. □

Definition 2.2.2. *We can define the model averaging weights for this comparison by:*

$$P_\alpha(M_1 \succ M_2) = \frac{\exp\left(W_\alpha(M_1, M_2)\right)}{1 + \exp\left(W_\alpha(M_1, M_2)\right)}$$

where $P(M_1 \succ M_2)$ is the probability that model M_1 is preferred to model M_2. Note that this definition provides $P(M_1 \succ M_2) = 1 - P(M_2 \succ M_1)$.

Proposition 2.2.3. $\exp\left(W_\alpha(M_1, M_2)\right)$ *is a discounted version of the posterior Bayes' Factor.*

Proof. Simple calculation provides:

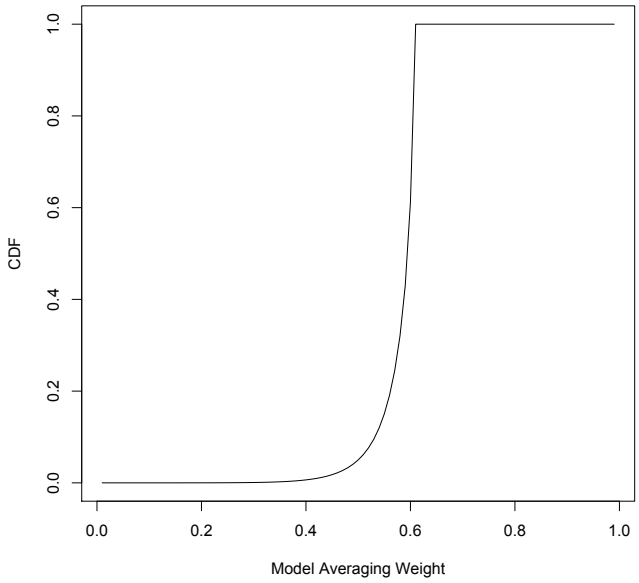
$$\exp\left(W_\alpha(M_1, M_2)\right) = \frac{f_1(\boldsymbol{Y}|\boldsymbol{Y}) \exp\left(-D(f_1 \parallel f_2)\right)}{f_2(\boldsymbol{Y}|\boldsymbol{Y}) \exp\left(-D(f_2 \parallel f_1)\right)}$$

□

Remark 2.2.2. *It is interesting to note that using the above model averaging weight for the test in Proposition 2.2.2 provides a good model averaging weight for M_1 when M_1 is true, dependent on the level of significance chosen. In fact, choosing a Type*

I error rate of 0.05 *provides* $\alpha \approx 1.548$ *and we can plot the cumulative distribution function of the model averaging weight (see Fig 2.2.2). The model averaging weight has an upper limit of about* 0.607.

Figure 2.1. CDF of model averaging weights for Remark 2.2.2



Proposition 2.2.4. *Suppose that* $M_1 : y_i|\boldsymbol{X}_i \overset{iid}{\sim} \mathcal{N}(\boldsymbol{X}_i\boldsymbol{\beta}_{\mathbf{1}i}, 1)$ *and that* $M_2 : y_i|\boldsymbol{X}_i, \boldsymbol{Z}_i \overset{iid}{\sim} \mathcal{N}(\boldsymbol{X}_i\boldsymbol{\beta}_{\mathbf{2}i} + \boldsymbol{Z}_i\boldsymbol{\gamma}_i, 1)$ *for* $i = 1, \ldots, n$ *and that the priors on all of the coefficients in each model is a constant. Then the test is can be made to have arbitrarily small Type I error rate (with for* $0.5 < \alpha < 2$*). In particular, as* $|\boldsymbol{\gamma}|$ *grows the probability of rejecting* M_1 *when it is true goes to* 0 *for fixed* $\alpha \in (.5, 2)$.

Proof. Let $\boldsymbol{X}$ be a matrix whose i-th row is $\boldsymbol{X}_i$. Define $\boldsymbol{Z}$ similarly. Without loss of generality, we can orthogonalize the columns of $\boldsymbol{Z}$ to the columns of $\boldsymbol{X}$. Let

$q = |\boldsymbol{\beta_k}|$, $r = |\boldsymbol{\gamma}|$, $\boldsymbol{Q} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$, and $\boldsymbol{R} = \boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T$. Suppose that we have gathered data $\boldsymbol{Y}$. We can compute the difference in K-L Divergences by letting $\boldsymbol{S} = \boldsymbol{I} - \frac{1}{2}\boldsymbol{Q}$, $\boldsymbol{T} = \boldsymbol{S} - \frac{1}{2}\boldsymbol{R}$, $\boldsymbol{\mu} = \boldsymbol{QY}$, and $\boldsymbol{\theta} = \boldsymbol{\mu} + \boldsymbol{RY}$. We get a difference of K-L Divergences which is

$$I(M_1, M_2) = r\log(2) - \frac{3r}{4} - \frac{1}{4}\boldsymbol{Y}^T\boldsymbol{RY}$$

Evaluating the log predictive densities at the data $\boldsymbol{Y}$ provides us with

$$\log\left(\frac{f_1(Y|Y)}{f_2(Y|Y)}\right) = r\frac{\log(2)}{2} - \frac{1}{2}\boldsymbol{Y}^T\boldsymbol{RY}$$

We get a discrimination measure of

$$W_\alpha(M_1, M_2) = \frac{3r\alpha}{4} - r\frac{\log(2)(2\alpha - 1)}{2} - \frac{2-\alpha}{4}\boldsymbol{Y}^T\boldsymbol{RY}$$

Whenever M_1 is true, we know that $\boldsymbol{Y}^T\boldsymbol{RY}$ has a chi-squared distribution with r degrees of freedom. Using properties of this distribution, we can choose α to set the Type I error rate. In particular, for any $\alpha > .5$, we know that $\frac{3*\alpha - 2\log(8)(2\alpha-1)}{2-\alpha} > 1$ and the probability that $W_\alpha > 0$ is increasing to 1 as r increases for any fixed $0.5 < \alpha < 2$. Whenever M_2 is the true model, we know that $\boldsymbol{Y}^T\boldsymbol{RY}$ has a non-central chi-squared distribution with r degrees of freedom and non-centrality parameter $\lambda_n = \boldsymbol{\gamma}^*\boldsymbol{Z}^T\boldsymbol{Z}\boldsymbol{\gamma}^*$ where $\boldsymbol{\gamma}^*$ is the actual value of $\boldsymbol{\gamma}$ that generated the data. In fact, $\lambda_n = n\lambda$ where $\lambda > 0$, it is easy to see that $P(W < 0)$ is increasing as n grows. □

Proposition 2.2.5. *Suppose that we have two linear models with known precision and design matrices* $\boldsymbol{X}_1 = \boldsymbol{X} + \boldsymbol{Z}_1$ *and* $\boldsymbol{X}_2 = \boldsymbol{X} + \boldsymbol{Z}_2$ *where* $\boldsymbol{Z}_i$ *is orthogonal to* $\boldsymbol{X}$

for $i = 1, 2$. Define $\boldsymbol{Q} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}$ and $\boldsymbol{R}_i = \boldsymbol{Z}_i(\boldsymbol{Z}_i^T\boldsymbol{Z}_i)^{-1}\boldsymbol{X}_i$ where $\boldsymbol{Q}$ and $\boldsymbol{R}_i$ are full rank and have ranks q and r_i $(i = 1, 2)$. Suppose we have collected data $\boldsymbol{Y}$ of size n. For any $\alpha \in (\frac{1}{2}, 2)$ if a model is correct, it is chosen with probability 1 as n grows. Additionally, If the two models have the same residual sum of squared errors, then the model with fewer parameters is chosen.

Proof. This is easily established by noting that

$$W_\alpha(M_1, M_2) = (r_2 - r_1)\left[\frac{3\alpha - 2\log(2)(2-\alpha)}{4}\right] + \frac{2-\alpha}{4}\boldsymbol{Y}^T(\boldsymbol{R}_1 - \boldsymbol{R}_2)\boldsymbol{Y}$$

Clearly the last two claims are true. There are two independent non-central chi-squared distributions in the formula for W_α ($\boldsymbol{Y}^T\boldsymbol{R}_i\boldsymbol{Y}$ for $i = 1, 2$). If one of the models is true, its non-centrality parameter will be the larger of the two non-centrality parameters and that model will be chosen with probability 1 as n grows. In fact, if neither model is the true model, then the model with the larger value of the non-centrality parameter will be chosen with probability 1 as n grows and if the models have the same non-centrality parameter, then the model with fewer parameters is chosen. □

Remark 2.2.3 (Varying α)**.** *As the examples showed, a fixed α provided a fixed Type I error rate asymptotically. In order to get consistency, one must allow $\alpha = \alpha(n) \to 2$ in an appropriate manner to balance Type I and Type II error rates. The examples discussed we can modify α to a function of n and obtain a test that is always asymptotically consistent when the number of parameters is fixed.*

Fix $\alpha_0 \in \left(\frac{1}{2}, 2\right)$*, define*

$$\alpha(n) = 2 - \frac{2 - \alpha_0}{a(n)}$$

Choosing $a(n)$ *as an increasing function of* n *such that*

$$\lim_{n\to\infty} a(n) = \infty \qquad \lim_{n\to\infty} \frac{a(n)}{n} = 0$$

provides us with the result that Type I error goes to 0 *in the nested case and that the growth of*

$$r_2 \times \frac{3\alpha - 2\log(2)(1 - 2\alpha)}{2 - \alpha}$$

is slower than n *and so Type II error also goes to* 0*. This also provides consistency in the case of non-nested linear models and so consistency can be achieved across all linear model comparisons with fixed parameter spaces (that is, not having a fixed set of covariates that one can include and not increasing this set of covariates as* $n \to \infty$*).*

2.2.2 Multiple Model Comparison

Although the criterion has been initially defined for the comparison of two models, it can be easily extended to handle multiple model comparison. Suppose that there are k models $M^1, \ldots, M^k$ and define the $W_\alpha(i)$ by

$$W_\alpha(i) = \left(\log\left(f_i(\boldsymbol{Y}|\boldsymbol{Y})\right) - \alpha \sum_{j=1}^{k} \int f_j(\boldsymbol{y}|\boldsymbol{Y}) \log\left(f_i(\boldsymbol{y}|\boldsymbol{Y})\right) \mathrm{d}\boldsymbol{y} \right)$$

Here it is assumed that all of the integrals in the definition are finite. The difference in two of these is given by

$$W_\alpha(i) - W_\alpha(j) = \log\left(\frac{f_i(\boldsymbol{Y}|\boldsymbol{Y})}{f_j(\boldsymbol{Y}|\boldsymbol{Y})}\right) - \alpha \sum_{\ell=1}^{k} \int f_\ell(\boldsymbol{y}|\boldsymbol{Y}) \log\left(\frac{f_i(\boldsymbol{y}|\boldsymbol{Y})}{f_j(\boldsymbol{y}|\boldsymbol{Y})}\right) \mathrm{d}\boldsymbol{y}$$

which has a direct interpretation as a correction to the posterior Bayes' Factor which takes into account the predicted posterior Bayes factor when assuming each model in the class of considered models is correct. The quantity

$$D(g||f;h) = \int \log\left(\frac{g(x)}{f(x)}\right) h(x)\mathrm{d}x = D(f||h) - D(g||h)$$

is a difference in K-L divergences and represents the difference in the amount of information gained over f versus that gained over g when h is true distribution. The difference in W_αs becomes

$$W_\alpha(i) - W_\alpha(j) = \log\left(\frac{f_i(\boldsymbol{Y}|\boldsymbol{Y})}{f_j(\boldsymbol{Y}|\boldsymbol{Y})}\right) - \alpha\sum_{\ell=1}^{k} D(f_i||f_j;f_\ell)$$

The decomposition of W_α into terms representing fit and complexity can be easily seen. Each model M_i has a posterior mode $\widehat{\theta}_i$. The model selection criterion then decomposes as

$$\begin{aligned} W_\alpha(i) =& \log\left(f_i(\boldsymbol{Y}|\boldsymbol{Y})\right) - \alpha\sum_{j=1}^{k}\int f_j(\boldsymbol{y}|\boldsymbol{Y})\log\left(f_i(\boldsymbol{y}|\widehat{\theta}_i)\right)\mathrm{d}\boldsymbol{y} \\ & - \alpha\sum_{j=1}^{k}\int f_j(\boldsymbol{y}|\boldsymbol{Y})\log\left(\frac{f_i(\boldsymbol{y}|\boldsymbol{Y})}{f_i(\boldsymbol{y}|\widehat{\theta}_i)}\right)\mathrm{d}\boldsymbol{y} \end{aligned}$$

The term

$$\alpha\sum_{j=1}^{k}\int f_j(\boldsymbol{y}|\boldsymbol{Y})\log\left(\frac{f_i(\boldsymbol{y}|\boldsymbol{Y})}{f_i(\boldsymbol{y}|\widehat{\theta}_i)}\right)\mathrm{d}\boldsymbol{y}$$

is a natural expression for the complexity of M_i. In comparison to the complexity term in the DIC

$$p_D = \int \log\left(\frac{f_i(\boldsymbol{Y}|\boldsymbol{\theta}_i)}{f_i(\boldsymbol{Y}|\boldsymbol{\theta}_i^*)}\right) p_i(\boldsymbol{\theta}_i|\boldsymbol{Y})\mathrm{d}\boldsymbol{\theta}_i$$

the term in $W_\alpha(i)$ expresses complexity in terms of the uncertainty in predicted values, and not the uncertainty in estimates of the parameter. This measure of complexity takes into account the entire class of models under consideration as it is the sum of integrals with respect to the predictive density of each model under consideration.

On the other hand, the term

$$\left(\log\left(pr_i(\boldsymbol{Y}|\boldsymbol{Y})\right) - \alpha \sum_{j=1}^{k} \int pr_j(\boldsymbol{y}|\boldsymbol{Y}) \log\left(f_i(\boldsymbol{y}|\widehat{\theta}_i)\right) \mathrm{d}\boldsymbol{y} \right)$$

is a natural term for model fit. It is the difference of the log Posterior Bayes' Factor and a sum of expectations of the log-likelihood evaluated at the posterior mode. This sum takes into account each model in the class considered, reflecting uncertainty about which model in the class should be assumed correct when considering the fit of a particular model. Of particular interest for W_α is the fact that it decomposes as terms that represent fit and complexity while being invariant to model focus, in contrast to other information criteria that penalize based on complexity. The key innovation is that W_α is not a model internal criterion, but a criterion that engages all models in the class of considered models while determining the fit and complexity of a particular model.

2.3 Predictively Modified Bayes' Factors

In contrast to the PPIC, where the criteria is defined in a model external manner using only predictive densities, we define a model internal selection criteria in the section using both predictive and marginal densities. The key is to combine the

insights of the fractional and intrinsic Bayes' Factors with predictive densities to obtain a large family of criteria that behave like Bayes' Factors but also provide the investigator with a certain amount of flexibility and control. For this section we have a single statistical model with the following densities (distributions if necessary):

$$\text{Prior Density of } \boldsymbol{\theta}: \ \pi(\boldsymbol{\theta})$$

$$\text{Sampling Density of } \boldsymbol{y}: \ f(\boldsymbol{y}|\boldsymbol{\theta})$$

$$\text{Marginal Density of } \boldsymbol{y}: \ m(\boldsymbol{y}) = \int f(\boldsymbol{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}$$

$$\text{Posterior Density of } \boldsymbol{\theta}: \ p(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{f(\boldsymbol{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{m(\boldsymbol{y})}$$

$$\text{Predictive Density of New Data } \boldsymbol{z}:$$

$$pr(\boldsymbol{z}|\boldsymbol{y}) = \int f(\boldsymbol{z}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{y})\mathrm{d}\boldsymbol{\theta}$$

We have formed the posterior predictive density under the assumption that $\boldsymbol{z} \perp\!\!\!\perp \boldsymbol{y}|\boldsymbol{\theta}$. This choice of predictive density is part of the formulation of the model and represents how one views the data generation process for future observations. We will call I the information we need to form this posterior predictive density and implicitly condition on it throughout the section. It is important to note that I also describes some information about $\boldsymbol{z}$, for example its length and any design considerations that need to go into creating an appropriate distribution, as well as the prior π used to form the posterior distribution.

Define the αth modified marginal to be

$$m_\alpha(\boldsymbol{y}) = \left(\int pr(\boldsymbol{y}|\boldsymbol{z})^{\alpha-1} pr(\boldsymbol{z}|\boldsymbol{y})\mathrm{d}\boldsymbol{z}\right)^{\frac{1}{\alpha-1}} = m(\boldsymbol{y}) \exp\left(D_\alpha(pr(\cdot|\boldsymbol{y})||m(\cdot))\right)$$

Before we begin with the general properties of this family of criteria, we will review the case $\alpha = 2$ and $\boldsymbol{z}$ is a minimal sample with some detail. This is precisely the method devised by Iwaki.

2.3.1 Iwaki's Expected Posterior Predicted Priors

In his 1997 paper, Iwaki considered taking prior $\pi(\boldsymbol{\theta})$ that are improper and proposed replacing this improper prior with a kind of posterior distribution that could be learned from the data. In order to avoid the issues with over fitting in Aitkin's posterior Bayes' factor, he suggested replacing the prior with

$$\tilde{\pi}(\boldsymbol{\theta}) = \int p(\boldsymbol{\theta}|\boldsymbol{z}) pr(\boldsymbol{z}|\boldsymbol{y}) \mathrm{d}\boldsymbol{z}$$

where z is a sample of the smallest number of observations such that $p(\boldsymbol{\theta}|\boldsymbol{z})$ is a proper density. Such a sample is what we mean by a minimal training sample. This modified prior was proposed because it is indeed proper (as is easily shown using Fubini's Theorem) and, though it is data dependent, it is in a somewhat weak way. Suppose that the original posterior is consistent at the true value $\boldsymbol{\theta}_0$ and that the parameter space is a complete and separable metric space. Further assume that under a minimal sample $f(\boldsymbol{z}|\boldsymbol{\theta})$ is a bounded function of $\boldsymbol{\theta}$ a.e. $pr(\boldsymbol{z}|\boldsymbol{y})$. Since $p(\boldsymbol{\theta}|\boldsymbol{y}) \xrightarrow{w} \delta_{\boldsymbol{\theta}_0}(\boldsymbol{\theta})$ we have

$$pr(\boldsymbol{z}|\boldsymbol{y}) = \int f(\boldsymbol{z}|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\boldsymbol{y}) \mathrm{d}\boldsymbol{\theta} \to f(\boldsymbol{z}|\boldsymbol{\theta}_0) \ a.e.$$

This suggests that $\tilde{\pi}$, which is $p(\boldsymbol{\theta}|\boldsymbol{z})$ integrated against $pr(\boldsymbol{z}|\boldsymbol{y})$, will at least exhibit the uncertainty induced by all possible minimal samples arising from the true process.

Defining $\tilde{\pi}(\boldsymbol{\theta}'|\boldsymbol{\theta}) = \int p(\boldsymbol{\theta}'|\boldsymbol{z}) f(\boldsymbol{z}|\boldsymbol{\theta}) \mathrm{d}\boldsymbol{z}$ and changing the order of integration using Fubini (making the necessary measurability assumptions), we can see that

$$\tilde{\pi}(\boldsymbol{\theta}') = \int \tilde{\pi}(\boldsymbol{\theta}'|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\boldsymbol{y}) \mathrm{d}\boldsymbol{\theta}$$

which shows that $\tilde{\pi}(\boldsymbol{\theta}'|\boldsymbol{\theta}) \to \tilde{\pi}(\boldsymbol{\theta}'|\boldsymbol{\theta}_0)$. These priors are a correction to the original prior which exhibits minimal learning while not conditioning on a particular minimal sample (or averaging over all possible minimal samples like in the IBF).

Since the prior obtained is proper and non-degenerate as $n \to \infty$, it leads to the same sort of asymptotic model selection behavior as any marginal obtained from a properly defined subjective prior. Beyond using this machine to create proper priors from improper priors, it does not seem unreasonable to use it in robustifying a proper prior. If a proper prior is specified poorly, then this method can create a prior which pulls that prior towards the distribution that the data suggests is true while not overpowering the information in the prior (unless a minimal sample would do so anyhow). We conclude this discussion of Iwaki's method, we present three examples of these priors.

Example 2.3.1 (Normal Mean)**.** *Suppose that* $y_i \overset{iid}{\sim} \mathcal{N}(\mu, 1)$ *and that the prior density is* $\pi(\mu) \propto 1$*. The posterior predictive distribution is* $\mathcal{N}\left(\overline{y}, \frac{n+1}{n}\right)$ *and* $\tilde{\pi}$ *is a Gaussian with mean* $\overline{y}$ *and standard deviation* $\sqrt{\frac{2n+1}{n}}$*. This example most clearly demonstrates the fact that* $\tilde{\pi}$ *converges to a distribution that is not degenerate as* $n \to \infty$*. In fact, if* $\overline{y} \to \mu_0$*, then this prior is* $\mathcal{N}(\mu_0, \sqrt{2})$*.*

Example 2.3.2 (Bernoulli Distribution)**.** *Suppose that y_i are iid Bernoulli with parameter θ. Suppose that the prior is a Beta(α, β) distribution and we have observed s successes and f failures. If we let z be two observations, then*

$$\tilde{\pi}(\theta) = \frac{\Gamma(a+b+2)\theta^{a-1}(1-\theta)^{b-1}}{\Gamma(a)\Gamma(b)} \times \left(\theta^2 \frac{(s+a+1)(s+a)}{(a+1)(a)} + 2\theta(1-\theta)\frac{(f+b)(s+a)}{(b)(a)} + (1-\theta)^2 \frac{(f+b+1)(f+b)}{(b+1)(b)} \right)$$

If we take $\alpha = \beta \to 0$ (the Haldane prior), we get $\tilde{\pi}$ as a mixture of three pieces. Two are point masses at 0 and 1 with weights $\frac{f(f+1)}{n(n+1)}$ and $\frac{s(s+1)}{n(n+1)}$, respectively. The third is a uniform distribution with weight $\frac{2sf}{n(n+1)}$.

Example 2.3.3 (Gamma Distribution)**.** *Suppose that y_i are iid Gamma(θ, b) with $b > 0$ fixed and $\pi(\theta) \propto \frac{1}{\theta}$. A minimal sample z has one observation. The posterior distribution is Gamma$(\sum y_i, nb)$ and*

$$\tilde{\pi}(\theta'|\theta) = \frac{(\theta')^{b-1}\theta^b \Gamma(2b)}{(\theta'+\theta)^{2b}\Gamma(b)^2}$$

Integration with respect θ is complicated, and $\tilde{\pi}$ is

$$\tilde{\pi}(\theta) = (\sum y_i)^b \theta^{b-1} \frac{\Gamma((n+1)b)}{\Gamma(b)^2\Gamma(nb)} \int_0^\infty \frac{u^{2b-1}}{(u+1)^{(n+1)b}} \exp\left(-u\sum y_i\theta\right) du$$

The integral in the formula converges and is similar to the confluent hypergeometric function (and similarly quite terrible to compute). Since the $\phi(x) = \exp((\,) - x)$ is a convex function, and $\frac{u^{2b-1}}{(u+1)^{(n+1}b}$ is integrable (n at least 2), we can

2.3.2 General Properties

First, we will try to characterize the maximum and minimum values of the modified marginals. Finding good upper bounds will be easy, but lower bounds will be more difficult. In addition, we will characterize the amount of information gain with the Rényi entropy is providing asymptotically, showing that it is bounded for an infinite class of criteria which will provide the same asymptotic behavior as a traditional marginal. Before we begin, we note that because the modified marginals can be defined purely in terms of predictive densities that the modified marginals do not exhibit issues of model focus and are not subject to the indeterminacy of the undefined constants that arise from improper priors.

Proposition 2.3.1. *Suppose that an MLE exists. Then* $m_\alpha(\boldsymbol{y}) \leq f(\boldsymbol{y}|\widehat{\boldsymbol{\theta}}_{\boldsymbol{y}})$ *so long as* $\boldsymbol{z}$ *is such that* $p(\boldsymbol{\theta}|\boldsymbol{z})$ *is a proper posterior.*

Proof. Suppose that $\alpha > 1$ and consider the integral $\int pr(\boldsymbol{y}|\boldsymbol{z})^{\alpha-1} pr(\boldsymbol{z}|\boldsymbol{y}) \mathrm{d}\boldsymbol{z}$. We can bound the term $pr(\boldsymbol{y}|\boldsymbol{z})$ by writing it as an integral and using the two assumptions.

$$pr(\boldsymbol{y}|\boldsymbol{z}) = \int f(\boldsymbol{y}|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\boldsymbol{z}) \mathrm{d}\boldsymbol{\theta} \leq \int f(\boldsymbol{y}|\widehat{\boldsymbol{\theta}}_{\boldsymbol{y}}) p(\boldsymbol{\theta}|\boldsymbol{z}) \mathrm{d}\boldsymbol{\theta} = f(\boldsymbol{y}|\widehat{\boldsymbol{\theta}}_{\boldsymbol{y}})$$

Thus for $\alpha > 1$, we have

$$m_\alpha(\boldsymbol{y}) = \left(\int pr(\boldsymbol{y}|\boldsymbol{z})^{\alpha-1} pr(\boldsymbol{z}|\boldsymbol{y}) \mathrm{d}\boldsymbol{z} \right)^{\frac{1}{\alpha-1}} \leq f(\boldsymbol{y}|\widehat{\boldsymbol{\theta}}_{\boldsymbol{y}}) \left(pr(\boldsymbol{z}|\boldsymbol{y}) \mathrm{d}\boldsymbol{z} \right)^{\frac{1}{\alpha-1}} = f(\boldsymbol{y}|\widehat{\boldsymbol{\theta}}_{\boldsymbol{y}})$$

To get the bound for general α, use the fact that the Rényi divergences are ordered, and so for $\alpha' \leq 1$ we have $m'_\alpha(\boldsymbol{y}) \leq m_\alpha(\boldsymbol{y}) \leq f(\boldsymbol{y}|\widehat{\boldsymbol{\theta}}_{\boldsymbol{y}})$. □

Proposition 2.3.2. *If π is a proper prior, then for any $\boldsymbol{z}$ and any $\alpha > 0$, we have $m_\alpha(\boldsymbol{y}) \geq m(\boldsymbol{y})$. In fact, the lower bound is characterized by the common support of the marginal and posterior predictive densities.*

Proof. The first statement follows trivially since the Rényi divergences are bounded below by 0. The second follows from the fact the $D_0(pr||m) = -\log\left(\int_A m\right)$ and so

$$m_\alpha(\boldsymbol{y}) \geq m(\boldsymbol{y} \times \frac{1}{\int_A m},$$

where $A = \{\boldsymbol{z} : pr(\boldsymbol{z}|\boldsymbol{y}) > 0\}$ is the support of the posterior predictive density. □

Remark 2.3.1. *This proposition, though entirely trivial to prove, shows the difficulty in finding a lower bound if we use an improper prior. The difficulty comes from the fact that $\int_{supp\ pr} m$ need not be finite for any sample $\boldsymbol{y}$ (we will see at least one example of this later). The problem is essentially this, once an improper measure is employed, it is impossible to determine that the Rényi divergences are bounded below. We will show an example where the change from existing for all $\alpha > 0$ and for only some α greater than a lower bound can be subtle and arise from small changes in model.*

Example 2.3.4 (Normal Means)**.** *Suppose that $y_i \overset{iid}{\sim} \mathcal{N}(\mu, 1)$ for $i = 1, \ldots, n$ with prior $\pi(\mu) \propto 1$. Suppose that $\boldsymbol{z}$ is a sample of size $n_0 \geq 1$. Then m_α exists for all $\alpha > 0$ and for all n_0. To show that m_α exists, it suffices to show that the $\boldsymbol{z}'\Sigma^{-1}\boldsymbol{z}$ term yields an integrable Gaussian from the quadratic form in the exponential of $pr(\boldsymbol{y}|\boldsymbol{z})^{\alpha-1}pr(\boldsymbol{z}|\boldsymbol{y})$. This quadratic form is provides*

$$\Sigma^{-1} = \boldsymbol{I}_{n_0} + \mathbf{1}_{n_0}(\mathbf{1}'_{n_0}\mathbf{1}_{n_0})^{-1}\mathbf{1}'_{n_0}\left(\frac{\alpha n}{n + n_0} - 1\right),$$

which is positive definite so long as $\frac{\alpha n}{n+n_0} > 0$*, which is trivially true for and* $\alpha > 0$*. Note that under this improper prior, the value* $\alpha = 0$ *does not give rise to a positive definite matrix.*

In contrast, suppose that $y_i \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$ *for* $i = 1, \ldots, n$ *with prior* $\pi(\mu, \sigma) \propto \sigma^{-1}$*. Suppose that* $\boldsymbol{z}$ *is a sample of size* $n_0 \geq 2$*. Then* m_α *exists for all* $\alpha > \frac{1}{n}$ *and for all* n_0*. To show that* m_α *exists, we must compute the general form, which turns out to be*

$$m_\alpha(\boldsymbol{y}) = \left(\frac{n_0}{n+n_0}\right)^{\frac{1}{2}} \frac{1}{(\sqrt{\pi})^n} \frac{1}{(\boldsymbol{y}^T(\boldsymbol{I}_n - \boldsymbol{1}_n(\boldsymbol{1}_n'\boldsymbol{1}_n)^{-1}\boldsymbol{1}_n')\boldsymbol{y})^{\frac{n}{2}}} \times \frac{\Gamma\left(\frac{n-1}{2}\right)^{\frac{1}{1-\alpha}} \Gamma\left(\frac{n_0-1}{2}\right)^{\frac{\alpha}{1-\alpha}}}{\Gamma\left(\frac{n+n_0-1}{2}\right)^{\frac{\alpha}{1-\alpha}}} \left(\frac{\Gamma\left(\frac{(n+n_0-1)\alpha}{2}\right)}{\Gamma\left(\frac{n\alpha-1}{2}\right)\Gamma\left(\frac{(n_0-1)\alpha}{2}\right)}\right)^{\frac{1}{1-\alpha}}$$

To see why $\alpha > \frac{1}{n}$*, consider the* $\Gamma\left(\frac{n\alpha-1}{2}\right)$ *in the denominator. Since the* Γ *function has a singularity at* 0*, this term makes the entire modified marginal* 0 *as* $\alpha \to \frac{1}{n}$ *from above. Of course, we could define this quantity for* α *for* $0 < \alpha < \frac{1}{n}$ *but it is not clear how that will behave with respect to the rest of the modified marginals (for example, we might lose the ordering property).*

As the next theorem shows, finding a lower bound for the modified marginal is not problematic for fixed $\alpha > 1$ when minimal samples are available. In order to do this, we need to make some assumptions about the posterior distributions and some information integrals which are relatively weak in the case of exchangeable observations. In fact, what we show is that the Rényi divergence converges to a stable value for fixed $\alpha > 0$ and $\boldsymbol{z}$ of a fixed sample size.

Theorem 2.3.1. *Suppose that the posterior distribution converges weakly to $\delta_{\boldsymbol{\theta}_0}$ where θ_0 is the true value of $\boldsymbol{\theta}$ and that we have the necessary measureability assumptions to apply Fubini's Theorem where necessary. Further suppose that $\boldsymbol{z}$ is a replicated sample with fixed sample size which is at least the minimal sample size. Assume that $f(\boldsymbol{z}|\boldsymbol{\theta})$ is bounded as a function of $\boldsymbol{\theta}$ and integrable with respect to $p(\boldsymbol{\theta}|\boldsymbol{y})$ for n large enough and that $D_\alpha(\boldsymbol{\theta}) = \exp\left(D_\alpha(f(\cdot|\boldsymbol{\theta})||m(\cdot))\right)$ is a bounded function of $\boldsymbol{\theta}$ and integrable with respect to $p(\boldsymbol{\theta}|\boldsymbol{y})$ for n large enough. Then*

$$\exp\left(D_\alpha(pr(\cdot|\boldsymbol{y})||m(\cdot))\right) \to D_\alpha(\boldsymbol{\theta}_0)$$

Proof. Let n be sufficiently large. First, use Fatou's Lemma, the assumption that $f(\boldsymbol{z}|\boldsymbol{\theta})$ is bounded, and that $p(\boldsymbol{\theta}|\boldsymbol{y}) \xrightarrow{w} \delta_{\boldsymbol{\theta}_0}$ to show that

$$\begin{aligned}
D_\alpha(\boldsymbol{\theta}_0)^{\alpha-1} &= \int \left(\frac{f(\boldsymbol{z}|\boldsymbol{\theta}_0)^\alpha}{m(\boldsymbol{z})}\right) \mathrm{d}\boldsymbol{z} \\
&= \int \left(\frac{\left(\lim \int f(\boldsymbol{z}|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\boldsymbol{y}) \mathrm{d}\boldsymbol{\theta}\right)^\alpha}{m(\boldsymbol{z})}\right) \mathrm{d}\boldsymbol{z} \\
&= \int \left(\frac{\left(\underline{\lim} \int f(\boldsymbol{z}|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\boldsymbol{y}) \mathrm{d}\boldsymbol{\theta}\right)^\alpha}{m(\boldsymbol{z})}\right) \mathrm{d}\boldsymbol{z} \\
&= \int \underline{\lim} \left(\frac{\left(\int f(\boldsymbol{z}|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\boldsymbol{y}) \mathrm{d}\boldsymbol{\theta}\right)^\alpha}{m(\boldsymbol{z})}\right) \mathrm{d}\boldsymbol{z} \\
&= \int \underline{\lim} \left(\frac{pr(\boldsymbol{z}|\boldsymbol{y})^\alpha}{m(\boldsymbol{z})}\right) \mathrm{d}\boldsymbol{z} \\
&\leq \underline{\lim} \int \left(\frac{pr(\boldsymbol{z}|\boldsymbol{y})^\alpha}{m(\boldsymbol{z})}\right) \mathrm{d}\boldsymbol{z} \\
&\leq \overline{\lim} \int \left(\frac{pr(\boldsymbol{z}|\boldsymbol{y})^\alpha}{m(\boldsymbol{z})}\right) \mathrm{d}\boldsymbol{z}
\end{aligned}$$

Now, use the assumption that $\alpha > 1$, Jensen's inequality, and Fubini's Theorem to show that

$$\begin{aligned}
\int \frac{pr(\boldsymbol{z}|\boldsymbol{y})^{\alpha}}{m(\boldsymbol{z})}\mathrm{d}\boldsymbol{z} &= \int \frac{\left(\int f(\boldsymbol{z}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{y})\mathrm{d}\boldsymbol{\theta}\right)^{\alpha}}{m(\boldsymbol{z})}\mathrm{d}\boldsymbol{z} \\
&\leq \int \frac{\int \left(f(\boldsymbol{z}|\boldsymbol{\theta})\right)^{\alpha} p(\boldsymbol{\theta}|\boldsymbol{y})\mathrm{d}\boldsymbol{\theta}}{m(\boldsymbol{z})}\mathrm{d}\boldsymbol{z} \\
&= \int\int \frac{\left(f(\boldsymbol{z}|\boldsymbol{\theta})\right)^{\alpha}}{m(\boldsymbol{z})} p(\boldsymbol{\theta}|\boldsymbol{y})\mathrm{d}\boldsymbol{\theta}\mathrm{d}\boldsymbol{z} \\
&= \int\int \frac{\left(f(\boldsymbol{z}|\boldsymbol{\theta})\right)^{\alpha}}{m(\boldsymbol{z})}\mathrm{d}\boldsymbol{z} p(\boldsymbol{\theta}|\boldsymbol{y})\mathrm{d}\boldsymbol{\theta} \\
&= \int D_{\alpha}(\boldsymbol{\theta})^{\alpha-1} p(\boldsymbol{\theta}|\boldsymbol{y})\mathrm{d}\boldsymbol{\theta}
\end{aligned}$$

The assumption that D_{α} is bounded and the weak convergence of the posterior implies that

$$\int D_{\alpha}(\boldsymbol{\theta})^{\alpha-1} p(\boldsymbol{\theta}|\boldsymbol{y})\mathrm{d}\boldsymbol{\theta} \to D_{\alpha}(\boldsymbol{\theta}_0)^{\alpha-1}$$

Thus, we can conclude that

$$\begin{aligned}
D_{\alpha}(\boldsymbol{\theta}_0)^{\alpha-1} &\leq \underline{\lim} \exp\left((\alpha-1)D_{\alpha}(pr(\cdot|\boldsymbol{y})||m(\cdot))\right) \\
&\leq \overline{\lim} \exp\left((\alpha-1)D_{\alpha}(pr(\cdot|\boldsymbol{y})||m(\cdot))\right) \\
&\leq D_{\alpha}(\boldsymbol{\theta}_0)^{\alpha-1}
\end{aligned}$$

and so

$$\lim \exp\left(D_{\alpha}(pr(\cdot|\boldsymbol{y})||m(\cdot))\right) = D_{\alpha}(\boldsymbol{\theta}_0)$$

□

As a corollary, we can conclude using the result from Iwaki's method that for any model, $\boldsymbol{z}$, and α satisfying the conditions of the theorem, we get behavior that

is asymptotically equivalent to using the Bayes' Factor from a proper prior. This is a rather strong statement. Under relatively mild conditions, an entire class of information criteria that allow the use of both proper and improper priors behaves asymptotically as classical Bayesian methods of model comparison. To see where the proof breaks down for $0 < \alpha \leq 1$, note that we cannot use Jensen's inequality (as $\phi(x) = x^\alpha$ is no longer convex) to get an appropriate upper bound for the integral. When minimal samples are available, there is hardly any reason to want to use $0 < \alpha \leq 1$, but when using fully replicated experiments we will have to use $\alpha \to 0^+$ in order to penalize complex models enough to get consistency. The proof of the theorem does show that we can get an upper bound for appropriate $0 < \alpha < 1$ since

$$D_\alpha(\boldsymbol{\theta}_0) \geq \overline{\lim} \left(\frac{1}{\int m(\boldsymbol{z})^{1-\alpha} pr(\boldsymbol{z}|\boldsymbol{y}) \mathrm{d}\boldsymbol{z}} \right)^{1-\alpha}.$$

We conclude this discussion of existence and asymptotic properties by showing that the positivity of D_α for fractional α and a fully replicated experiment is equivalent to showing the positivity for a minimal sample when a minimal example exists.

Theorem 2.3.2. *Suppose that $0 < \alpha < 1$ and that $\boldsymbol{z}$ is a fully replicated experiment. Further assume that all of the z_is and y_is are exchangeable. Let $\boldsymbol{z}_0$ be any minimal sample from $\boldsymbol{z}$ and $\boldsymbol{z}_1$ be the data comprising of the rest of the z_is. Assume that order of integration with respect to the z_i can be arbitrarily reordered without changing the value of the integral. Then*

$$D_\alpha(pr(\boldsymbol{z}|\boldsymbol{y})||m(\boldsymbol{z})) \leq D_\alpha(pr(\boldsymbol{z}_0|\boldsymbol{y})||m(\boldsymbol{z}_0)).$$

Proof. First, notice that $\alpha < 1$ implies that

$$D_\alpha(pr(\boldsymbol{z}|\boldsymbol{y})||m(\boldsymbol{z})) = \left(\frac{1}{\int m(\boldsymbol{z})^{1-\alpha} pr(\boldsymbol{z}|\boldsymbol{y})^\alpha \mathrm{d}\boldsymbol{z}}\right)^{\frac{1}{1-\alpha}}$$

and so finding a lower bound for $D_\alpha(pr(\boldsymbol{z}|\boldsymbol{y}||m(\boldsymbol{z}))$ is equivalent to finding an upper bound for $\int m(\boldsymbol{z})^{1-\alpha} pr(\boldsymbol{z}|\boldsymbol{y}^\alpha \mathrm{d}\boldsymbol{z}$. Now, use the assumption that we can arbitrarily order the integration to get

$$\int\int m(\boldsymbol{z}_1|\boldsymbol{z}_0)^{1-\alpha} pr(\boldsymbol{z}_1|\boldsymbol{z}_0, \boldsymbol{y})^\alpha \mathrm{d}\boldsymbol{z}_1 m(\boldsymbol{z}_0)^{1-\alpha} pr(\boldsymbol{z}_0|\boldsymbol{y})^\alpha \mathrm{d}\boldsymbol{z}_0.$$

The proof will be completed if we can show that $\int m(\boldsymbol{z}_1|\boldsymbol{z}_0)^{1-\alpha} pr(\boldsymbol{z}_1|\boldsymbol{z}_0, \boldsymbol{y})^\alpha \mathrm{d}\boldsymbol{z}_1 \geq 1$. Since $0 < \alpha < 1$, we can use the fact the $\phi(x) = x^\alpha$ is concave and the fact that $m(\boldsymbol{z}_1|\boldsymbol{z}_0)$ is a proper distribution to get

$$\begin{aligned}
\int m(\boldsymbol{z}_1|\boldsymbol{z}_0)^{1-\alpha} pr(\boldsymbol{z}_1|\boldsymbol{z}_0, \boldsymbol{y})^\alpha \mathrm{d}\boldsymbol{z}_1 &= \int m(\boldsymbol{z}_1|\boldsymbol{z}_0) \left(\frac{pr(\boldsymbol{z}_1|\boldsymbol{z}_0, \boldsymbol{y})}{m(\boldsymbol{z}_1|\boldsymbol{z}_0)}\right)^\alpha \mathrm{d}\boldsymbol{z}_1 \\
&\geq \left(\int m(\boldsymbol{z}_1|\boldsymbol{z}_0) \frac{pr(\boldsymbol{z}_1|\boldsymbol{z}_0, \boldsymbol{y})}{m(\boldsymbol{z}_1|\boldsymbol{z}_0)} \mathrm{d}\boldsymbol{z}_1\right)^\alpha \\
&= \left(\int pr(\boldsymbol{z}_1|\boldsymbol{z}_0, \boldsymbol{y}) \mathrm{d}\boldsymbol{z}_1\right)^\alpha \\
&= 1
\end{aligned}$$

□

2.3.3 Choice of α

Though the choice of α is effectively arbitrary for large samples so long as α is fixed and the replicated data has a fixed sample size, the choice of α can have a large impact on the criteria for small sample sizes. The choice of α should be made

to achieve some small sample objective, such as obtaining a certain expectation or Type I error rate. Since the Rényi divergences are ordered, increasing α increases the amount of additional information one is providing to a model for comparison. This suggests that smaller values of α should favor "smaller" models while larger values of α should favor larger models. In order to exhibit this, we analyze a simple class of models where the Bayes' Factor using modified marginals has a known distribution when the nested model is true, the class of linear models with known precision.

Suppose that model M_i is that $\boldsymbol{y}|\boldsymbol{x}_i, \boldsymbol{\beta}_i \sim \mathcal{N}(\boldsymbol{x}_i\boldsymbol{\beta}_i, \boldsymbol{I}_n)$ with $\pi_i(\boldsymbol{\beta}_i) \propto c_i$. The modified marginal for a given model is

$$m_\alpha(\boldsymbol{y}|M_i) = \left(\left(\frac{1}{\alpha}\right)^{\frac{1}{\alpha-1}} \frac{q_i}{n+q_i}\right)^{\frac{p_i}{2}} \left(\frac{1}{2\pi}\right)^{\frac{n}{2}} \exp\left(-\frac{1}{2}\boldsymbol{y}^T(\boldsymbol{I}_n - \boldsymbol{H}_i)\boldsymbol{y}\right),$$

where the replicated data $\boldsymbol{z}_i$ is taken to have dimension q_i design matrix $\boldsymbol{u}_i$ with $n\boldsymbol{u}_i^T\boldsymbol{u}_i = q_i\boldsymbol{x}_i^T\boldsymbol{x}_i$, $\boldsymbol{H}_i = \boldsymbol{x}_i(\boldsymbol{x}_i^T\boldsymbol{x}_i)^{-1}\boldsymbol{x_i}^T$, and p_i is the dimension of $\boldsymbol{\beta}_i$. Assuming that we have a finite set of models so that $p_{max} = \max_i p_i$ and choose $q_i = q = p_{max}$, we can investigate the properties of ratios of modified marginals (the modified Bayes' Factors) for a minimal sample ($n = p_{max}$). For minimal samples, we get modified Bayes' Factors of the form

$$MoBF_\alpha(i,j) = \left(\left(\frac{1}{\alpha}\right)^{\frac{1}{\alpha-1}} \frac{1}{2}\right)^{\frac{p_j-p_i}{2}} \exp\left(-\frac{1}{2}\boldsymbol{y}^T(\boldsymbol{H}_j - \boldsymbol{H}_i)\boldsymbol{y}\right).$$

We can now select an α in order to satisy some desirable minimal sample criterion. Consider the case of nested hypotheses $M_i \subset M_j$, then the statistic $\boldsymbol{y}^T(\boldsymbol{H}_j -$

$\boldsymbol{H}_i)\boldsymbol{y}$ has a $\chi^2_{p_j-p_i}$ distribution when M_i is the true model and we can determine $P(Type\ I\ Error) = P(MoBF_\alpha(i,j) < 1 | M_i\ is\ true)$ by considering

$$P\left(\chi^2_{p_j-p_i} > (p_j - p_i)\left(log(2) + \frac{1}{\alpha - 1}\log(\alpha)\right)\right).$$

In particular, $\frac{1}{\alpha-1}\log(\alpha)$ is a decreasing function of α which is equal to $1 - \log(2)$ when $\alpha \approx 7.6166$. Choosing any $0 < alpha < 7.6166$ will provide a Type I Error rate which is less than .5, increasing to .5 as $p_j - p_i$ increases for $\alpha = 7.6166$. It is clear that the Type I Error rate diminishes as $n \to \infty$, and so choosing an α to control it for minimal samples provides a means of controlling it for all possible samples. In addition to controlling the Type I Error rate, we might ask that the expected value of the logarithm of the MoBF to be a certain value. It is natural to ask that its expected value be 0 when a nested hypothesis is true. It is easy to see that this occurs at the value $\alpha = 7.6166$.

When considering more complicated models, one can take advantage of MCMC techniques to compute expected Type I Error rates when the nested hypothesis is true. Generating minimal samples from the posterior predictive density of the nested model can provide an estimate of the density of the log of the modified Bayes' Factor for minimal samples, which can then be used to calibrate the value of α.

2.3.4 Analytical Examples

The first example we present is the fact that any fractional α provides a $D_\alpha > 0$ which exists for any replicated sample size. In order to do this, we simply show it for a minimal sample z, which has size 1.

Example 2.3.5 (Gamma Distribution and $0 < \alpha < 1$)**.** *Suppose that y_i and z are iid Gamma(θ, b) for $b > 0$ fixed and that $\pi(\theta) \propto \theta^{-1}$ and that $n \geq 1$. Then*

$$pr(z|\boldsymbol{y}) = \frac{\Gamma((n+1)b)z^{b-1}\left(\sum y_i\right)^{nb}}{\Gamma(nb)\Gamma(b)\left(z+\sum y_i\right)^{(n+1)b}}$$
$$pr(\boldsymbol{y}|z) = \frac{\Gamma((n+1)b)z^{b}\left(\prod y_i\right)^{b-1}}{\Gamma(b)^{n+1}\left(z+\sum y_i\right)^{(n+1)b}}$$

We get

$$\begin{aligned}
\int_0^\infty pr(\boldsymbol{y}|z)^{\alpha-1}pr(z|\boldsymbol{y})dz &= \int_0^\infty \left(\frac{\Gamma((n+1)b)z^{b}\left(\prod y_i\right)^{b-1}}{\Gamma(b)^{n+1}\left(z+\sum y_i\right)^{(n+1)b}}\right)^{\alpha-1} \\
&\qquad \times \frac{\Gamma((n+1)b)z^{b-1}\left(\sum y_i\right)^{nb}}{\Gamma(nb)\Gamma(b)\left(z+\sum y_i\right)^{(n+1)b}}dz \\
&= C(\boldsymbol{y},n,b,\alpha)\int_0^\infty \frac{z^{b\alpha-1}}{\left(z+\sum y_i\right)^{(n+1)b\alpha}}dz \\
&= \tilde{C}(\boldsymbol{y},n,b,\alpha)\int_0^\infty \frac{\tilde{z}^{b\alpha-1}}{\left(\tilde{z}+1\right)^{(n+1)b\alpha}}d\tilde{z} \\
&= \tilde{C}(\boldsymbol{y},n,b,\alpha)\int_0^1 u^{b\alpha-1}(1-u)^{nb\alpha-1}du \\
&= \tilde{C}(\boldsymbol{y},n,b,\alpha)B(b\alpha,nb\alpha)
\end{aligned}$$

where we have used the change of variables $\tilde{z} = \frac{z}{\sum y_i}$ and $u = \frac{\tilde{z}}{\tilde{z}+1}$ and B is the beta function. Therefore, for any replicated sample size and any $0 < \alpha < 1$ we can conclude that $D_\alpha > 0$.

The next example is one that pushes the understanding of improper priors and shows that in small samples on can get results from this method that might be less than satisfactory, even for simple problems. The reason is that the example is discrete and so one can get observations which are maximally consistent with points on the boundary of the parameter space. For this example, we also produce the results for the fractional marginal of O'Hagan and a particular choice of expected posterior prior. This comparison shows the added data dependence of the modified marginal on the data over the EPP method which is fully Bayesian.

Example 2.3.6 (Bernoulli Data with Haldane Prior)**.** *Suppose that y_i and z_j are iid Bernoulli(θ) for $i = 1, \ldots, n$ and $j = 1, 2$ and that the prior is $\pi(\theta) \propto \theta^{-1}(1-\theta)^{-1}$. We will treat the prior as though it is taken as the limit as $a \to 0$ of a proper Beta(a, a) prior after all computations have been done. Assume that we have observed s successes and f failures.*

First, consider the fractional method, where we take $b = \frac{1}{n}$, then

$$\int f(\boldsymbol{y}|\theta)^b \pi(\theta)\, d\theta = B\left(\frac{s}{n}, \frac{f}{n}\right)\binom{n}{s}^{\frac{1}{n}}$$

and so the fractional marginal is

$$\binom{n}{s}^{1-\frac{1}{n}} \frac{B(s, f)}{B\left(\frac{s}{n}, \frac{f}{n}\right)},$$

where B is the beta function. If we have observed $s = 0$ or $f = 0$, the fractional marginal is 1, but when we have $0 < \theta_0 < 1$ the probability of observing $s = 0$ or $f = 0$ goes to 0 and the fractional marginal is asymptotically a constant times $\frac{n}{sf}$

Second, consider using an EPP with $m^*(\boldsymbol{z})$ arising from a Bernoulli(.5) distribution. The expected posterior prior is a mixture of three pieces, point masses at 0 and 1 with weight .25 and a uniform piece with weight .5. If $s = 0$ or $f = 0$, this provides a marginal distribution

$$m(y) = \begin{cases} \frac{1}{4} + \frac{1}{2(n+1)} & f = 0 \text{ or } s = 0 \\ \frac{1}{2(n+1)} & 0 < s < n \end{cases}$$

However, the modified marginal exhibits some different behavior. For $\alpha = 2$, we have

$$m_2(y) = \begin{cases} 1 & f = 0 \text{ or } s = 0 \\ \frac{2sf}{n(n+1)^2} & 0 < s < n \end{cases}$$

If we have observed only successes or failures, then the method provides weight only to the appropriate point mass. However, if $0 < \theta_0 < 1$, then we have comparable asymptotic behavior to both the fractional and intrinsic methods. In fact, all three are asymptotically equivalent (up to a constant) as using a uniform prior whenever the true chance of success is neither 0 nor 1.

The final analytical example that we include is an example where minimal samples are available but force one to make some design choices. In this example, it might be more reasonable to consider a fully replicated experiment and use a fractional power. The modifies marginal will not exist for some set of α near 0, but this set will shrink as the sample size grows, and so we can handicap more complex models in an appropriate manner to get consistency. We present the modified marginal across the

spectrum of possible replicated samples and values of α in order to see what possible behaviors the modified marginal has.

Example 2.3.7 (Linear Model with Unknown Variance)**.** *Assume that* $\boldsymbol{Y}|\boldsymbol{\beta}, s, \boldsymbol{X}, \boldsymbol{u} \sim \mathcal{N}_n(\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{I}_n s^{-1})$ *and* $\pi((\beta), s) = cs^{-1}$ *where the length of* $\boldsymbol{\beta}$ *is* p*. Further assume that* $\boldsymbol{z}|\boldsymbol{Y}, \boldsymbol{\beta}, s, \boldsymbol{X}, \boldsymbol{u} \sim \mathcal{N}_q(\boldsymbol{u}\boldsymbol{\beta}, \boldsymbol{I}_q s^{-1})$ *where* $\boldsymbol{u}$ *is taken to be any* $q \times p$ *(where* $q > p$*) matrix such that* $n\boldsymbol{u}^T\boldsymbol{u} = q\boldsymbol{X}^T\boldsymbol{X}$*. One obtains the following distributions*

$$
\begin{aligned}
m(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{u}) &= c\frac{\Gamma\left(\frac{n-p}{2}\right)}{(\sqrt{\pi})^{n-p}|\boldsymbol{X}^T\boldsymbol{X}|^{\frac{1}{2}}}\left(\frac{1}{\boldsymbol{Y}^T(\boldsymbol{I_n} - \boldsymbol{H})\boldsymbol{Y}}\right)^{\frac{n-p}{2}} \\
m(\boldsymbol{z}|\boldsymbol{X}, \boldsymbol{u}) &= c\frac{\Gamma\left(\frac{q-p}{2}\right)}{(\sqrt{\pi})^{q-p}|\boldsymbol{u}^T\boldsymbol{u}|^{\frac{1}{2}}}\left(\frac{1}{\boldsymbol{z}^T(\boldsymbol{I_q} - \boldsymbol{K})\boldsymbol{z}}\right)^{\frac{q-p}{2}} \\
\boldsymbol{z}|\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{u} &\sim \mathcal{MVT}_q\left(n-p, \boldsymbol{u}\widehat{\boldsymbol{\beta}}_{\boldsymbol{Y}}, \left(\boldsymbol{I}_q + \frac{q}{n}\boldsymbol{K}\right)\frac{\boldsymbol{Y}^T(\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{Y}}{n-p}\right)
\end{aligned}
$$

where $\boldsymbol{H}, \boldsymbol{K}$ *are the standard hat matrices for* $\boldsymbol{X}$ *and* $\boldsymbol{u}$*, respectively. One obtains the fractional modified marginal*

$$
\begin{aligned}
m^{\alpha,I}(\boldsymbol{Y}|\boldsymbol{X}) = &\left(\frac{q}{n+q}\right)^{\frac{p}{2}}\frac{1}{(\pi\boldsymbol{Y}^T(\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{Y})^{\frac{n}{2}}} \\
&\times\left(\frac{\Gamma\left(\frac{n+q-p}{2}\right)}{\Gamma\left(\frac{q-p}{2}\right)}\right)^{\frac{\alpha}{\alpha-1}}\left(\frac{\Gamma\left(\frac{n\alpha-p}{2}\right)\Gamma\left(\frac{(q-p)\alpha}{2}\right)}{\Gamma\left(\frac{(n+q-p)\alpha}{2}\right)\Gamma\left(\frac{n-p}{2}\right)}\right)^{\frac{1}{\alpha-1}}
\end{aligned}
$$

It is easy to see that the modified marginal exists only for $q > p$ *and* $\alpha > \frac{p}{n}$*. There are a few cases of interest to be considered, each of which can lead to consistent model selection. However, these cases also point out important differences in the values obtained when using a fully replicated experiment and* $\alpha \to 0$ *and those obtained for fixed* α *and* q*.*

Case (α, p, q fixed, $n \to \infty$)**.** *The modified marginal is asymptotically a constant,* $C(\alpha, q, p)$*, times*

$$\left(\frac{1}{n}\right)^{\frac{p-1}{2}} \left(\frac{2e\pi \boldsymbol{Y}^T(\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{Y}}{n}\right)^{-\frac{n}{2}}$$

For two given linear models with unknown variance, the ratio of the modified marginals is then a constant (depending on $\alpha_1, p_1, q_1, \alpha_2, p_2, q_2$*) times the quantity*

$$\left(\frac{1}{n}\right)^{\frac{p_1-p_2}{2}} \left(\frac{Y^T(\boldsymbol{I}_n - \boldsymbol{H}_2)\boldsymbol{Y}}{Y^T(\boldsymbol{I}_n - \boldsymbol{H}_1)\boldsymbol{Y}}\right)^{\frac{n}{2}}$$

which are the operational terms from a proper Bayes' Factor between two such models.

Case (p, q fixed, $\alpha = \frac{a}{n}$, $a > p$ fixed, $n \to \infty$)**.** *The modified marginal is asymptotically a constant,* $C(a, q, p)$*, times*

$$\left(\frac{1}{n}\right)^{p+\frac{a+1}{2}} \left(\frac{2e\pi \boldsymbol{Y}^T(\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{Y}}{n}\right)^{-\frac{n}{2}}$$

For two given linear models with unknown variance, the ratio of the modified marginals is then a constant (depending on $a_1, p_1, q_1, a_2, p_2, q_2$*) times the quantity*

$$\left(\frac{1}{n}\right)^{p_1-p_2+\frac{a_1-a_2}{2}} \left(\frac{Y^T(\boldsymbol{I}_n - \boldsymbol{H}_2)\boldsymbol{Y}}{Y^T(\boldsymbol{I}_n - \boldsymbol{H}_1)\boldsymbol{Y}}\right)^{\frac{n}{2}}$$

in order to maintain the same consistency results as are obtained by a proper Bayes' Factor, one needs to ensure that the exponent on $\frac{1}{n}$ *behaves appropriately. In particular, if* $p_1 > p_2$ *then one needs to ensure that* $a_2 - a_1 < 2(p_1 - p_2)$*.*

Case (p fixed, $\alpha = \frac{a}{n}$, $q = bn$, $a > p$ fixed, b fixed, $n \to \infty$)**.** *Though the constants change (as they depend now on* p, a, b*), the rest of the asymptotic form is similar to that in the last case. However, the power on* $\frac{1}{n}$ *is* $\frac{p+a-1}{2}$ *as opposed to* $p + \frac{a+1}{2}$*.*

We can clearly see that the asymptotic form once again relies on both p and a. To obtain consistency when comparing two such models with $p_1 > p_2$, one must maintain $a_2 - a_1 < p_1 - p_2$, which can be easily acheived by either choosing $a_1 = a_2$ or $a_i = p_i + 1$.

As is easily evidenced, using criteria between the cases to compare multiple models can easily lead to an inconsistent selection crtierion. For fixed q, any fixed α for each model leads to consistency. However, care must be taken when allowing q to grow or α to shrink and mixing criteria across the three cases is problematic.

2.3.5 Computational Examples

We present two computation examples where the difficulties in computation are overcome in two ways. The first is through the clever use of $\alpha = 2$ and a modification of the method to account for latent variables. The second is an example where the integral in $\boldsymbol{z}$ cannot be carried out directly, but both the marginals and posterior predictive densities are available in closed form.

Example 2.3.8 (Logistic Regression)**.** *One class of models that is particularly difficult for using this method is the analysis of dichotomous and polychotomous data. When analyzing a generalized linear model, the notion of a minimal sample and the calculation of marginal probabilities for outcomes is difficult and makes the computation of the modified marginal computationally burdensome. Drawing on recent work of*

[47]we can simplify computation by modifying the replicated data. Consider the probit model of [48]where we have the following hierarchical specification for $i = 1, \ldots, n$.

$$y_i|v_i, \boldsymbol{\beta}, \boldsymbol{x}_i = \begin{cases} 1 & \text{if } v_i > 0 \\ 0 & \text{if } v_i \leq 0 \end{cases}$$

$$v_i|\boldsymbol{\beta}, \boldsymbol{x}_i \sim \mathcal{N}(\boldsymbol{x}_i\beta, 1)$$

$$\boldsymbol{\beta}|\boldsymbol{x}_i \sim 1$$

where the dimension of $\boldsymbol{\beta}$ *is taken to be* p *and we assume that* $\boldsymbol{x}^T\boldsymbol{x}$ *is full rank and the covariates are assumed to be continuous.*

The idea is to ease the computation of the marginal distribution of replicated data by looking at the predictive densities of the latent variables as opposed to the predictive distributions of new observations. Essentially, we view the binary outcome as a loss of information from the underlying latent state and use the latent state build the information criterion. The really nice thing about this framework is that the marginal distribution of the latent state can be easily computed and a minimal sample is easy to consider. Let v_j^{rep} *be latent states for* $j = 1, \ldots, q$ *which arise from design a matrix* $\boldsymbol{x}^{rep}$ *where we restrict* $n(\boldsymbol{x}^{rep})^T\boldsymbol{x}^{rep} = q\boldsymbol{x}^T\boldsymbol{x}$. *The modified marginal that we will use is* $\left(\int pr(\boldsymbol{y}|\boldsymbol{v}^{rep})^{\alpha-1} pr(\boldsymbol{v}^{rep}|\boldsymbol{y})\, d\boldsymbol{v}^{rep}\right)^{\frac{1}{\alpha-1}}$ *where* $pr(\boldsymbol{v}^{rep}|\boldsymbol{y}) = \int f(\boldsymbol{v}^{rep}|\boldsymbol{\beta}) p(\boldsymbol{\beta}|\boldsymbol{y})\, d\boldsymbol{\beta}$. *This is equivalent to using the Rényi divergence*

$$D_\alpha = \left(\int \left(\frac{pr(\boldsymbol{v}^{rep}|\boldsymbol{y})}{m(\boldsymbol{v}^{rep})} \right)^{\alpha-1} pr(\boldsymbol{v}^{rep}|\boldsymbol{y})\, d\boldsymbol{v}^{rep} \right)^{\frac{1}{\alpha-1}}.$$

The marginal of a minimal sample is incredibly easy to compute under the unform prior and turns out to be a constant $c|(\boldsymbol{x}^{rep})^T\boldsymbol{x}^{rep}|^{-.5}$*, which will greatly facilitate computational aspects of the algorithm. Note that this is a minimal sample for the latent space and that it is not a minimal sample if we had to further generate* $\boldsymbol{y}^{rep}$*. Given the improper prior, it is hard to determine what sort of minimal sample* $\boldsymbol{y}^{rep}$ *one needs to obtain a finite marginal and the computation of such a marginal is somewhat arduous. In addition, using* $\alpha = 2$ *further reduces the computational burden since the order of integration in* $\boldsymbol{\beta}$ *and* $\boldsymbol{v}^{rep}$ *can be exchanged leading to an estimator of the divergence of*

$$\exp(D_2) \approx \frac{1}{c}|(\boldsymbol{x}^{rep})^T\boldsymbol{x}^{rep}|^{.5}\frac{1}{N^2}\sum_{i,j} f((\boldsymbol{v}^{rep})^{(i)}|\boldsymbol{\beta}^{(j)})$$

The final quantity needed to be estimated is the marginal for $\boldsymbol{y}$*, which can be done in the fashion of [49] using data augmentation*

$$m(\boldsymbol{y}) = \frac{f(\boldsymbol{y}|\boldsymbol{\beta}^*)c}{p(\boldsymbol{\beta}^*|\boldsymbol{y})} \text{ where } p(\boldsymbol{\beta}^*|\boldsymbol{y}) \approx \sum_j p(\boldsymbol{\beta}^*|\boldsymbol{v}^{(j)}).$$

Before showing a numerical example, consider the criterion asymptotically. Using consistency of the posterior, we can provide an estimator of the criterion by simply substiting the predictive density with the sampling density evaluated at the true value of the parameter,

$$\exp(D_2) \approx \left(\int pr(\boldsymbol{v}^{rep}|\boldsymbol{\beta}_0)^2 d\boldsymbol{v}^{rep}\right)\frac{1}{c}|(\boldsymbol{x}^{rep})^T\boldsymbol{x}^{rep}|^{.5} = \frac{|(\boldsymbol{x}^{rep})^T\boldsymbol{x}^{rep}|^{.5}}{c(4\pi)^{\frac{p}{2}}},$$

which does not depend on the value of $\boldsymbol{\beta}_0$. *Additionally, the MCMC calculation of* m_2 *is quite easy and can be done with the following steps repeated* N *times beginning with a base point* $\boldsymbol{\beta}^{(0)}$*:*

1. *Sample* $v_i^{(k)}$ *from* $v_i|\boldsymbol{\beta}^{(k-1)}, y_i$, *which is a truncated normal for* $i = 1, \ldots, n$.

2. *Sample* $\boldsymbol{\beta}^{(k)}$ *from* $\boldsymbol{\beta}|\boldsymbol{v}^{(k)}$, *which is multivariate normal.*

3. *Sample* $(\boldsymbol{v}^{rep})^{(k)}$ *from* $\boldsymbol{v}|\boldsymbol{\beta}^{(k)}$, *which is multivariate normal (in fact independent).*

4. *Sample* $(\boldsymbol{\beta}^{rep})^{(k)}$ *from* $\boldsymbol{\beta}|(\boldsymbol{v}^{rep})^{(k)}$

5. *Evaluate* $f(\boldsymbol{y}|(\boldsymbol{\beta}^{rep})^{(k)})$.

The estimate for m_2 *is then*

$$m_2 \approx \frac{1}{N}\sum_{k=1}^{N} f(\boldsymbol{y}|(\boldsymbol{\beta}^{rep})^{(k)}).$$

We exhibit this analysis using the nodal data from [49]

Since the data contains both continuous and categorical variables, we can choose a sample of size q *that respects the data structure and minimizes the Frobenius norm of* $q\boldsymbol{x}^T\boldsymbol{x} - n\boldsymbol{x}^{rep})^T\boldsymbol{x}^{rep}$. *Given the desire to accurately represent all of the possibilities of values for the categorical variables, we choose the sample to have* $q = 8$ *observations. The particular choice of minimal sample we used is presented in 2.3.8.*

A comparison of the log maximized likelihoods, marginals, and m_2 *is presented in table 2.3.8 where results are presented for both the proper prior used in [49]which is independent normal with all means being* .75 *and standard deviations being* 5. *There*

Table 2.1
Nodal Data from Chib 1995

	Y	X_1	X_2	X_3	X_4	X_5		Y	X_1	X_2	X_3	X_4	X_5
1	0	66	0.48	0	0	0	28	0	68	0.56	0	0	0
2	0	66	0.50	0	0	0	29	0	56	0.52	0	0	0
3	0	58	0.50	0	0	0	30	0	60	0.49	0	0	0
4	0	65	0.46	1	0	0	31	0	60	0.62	1	0	0
5	1	50	0.56	0	0	1	32	0	49	0.55	1	0	0
6	0	61	0.62	0	0	0	33	0	58	0.71	0	0	0
7	0	51	0.65	0	0	0	34	1	67	0.67	1	0	1
8	0	67	0.47	0	0	1	35	0	51	0.49	0	0	0
9	0	56	0.50	0	0	1	36	0	60	0.78	0	0	0
10	0	52	0.83	0	0	0	37	0	56	0.98	0	0	0
11	0	67	0.52	0	0	0	38	0	63	0.75	0	0	0
12	1	59	0.99	0	0	1	39	0	64	1.87	0	0	0
13	1	61	1.36	1	0	0	40	1	56	0.82	0	0	0
14	0	64	0.40	0	1	1	41	0	61	0.50	0	1	0
15	0	64	0.50	0	1	1	42	0	63	0.40	0	1	0
16	0	52	0.55	0	1	1	43	0	66	0.59	0	1	1
17	1	58	0.48	1	1	0	44	1	57	0.51	1	1	1
18	1	65	0.49	0	1	0	45	0	65	0.48	0	1	1
19	0	59	0.63	1	1	1	46	0	61	1.02	0	1	0
20	0	53	0.76	0	1	0	47	0	67	0.95	0	1	0
21	0	53	0.66	0	1	1	48	1	65	0.84	1	1	1
22	1	50	0.81	1	1	1	49	1	60	0.76	1	1	1
23	1	45	0.70	0	1	1	50	1	56	0.78	1	1	1
24	1	46	0.70	0	1	0	51	1	67	0.67	0	1	0
25	1	63	0.82	0	1	0	52	1	57	0.67	0	1	1
26	1	51	0.72	1	1	0	53	1	64	0.89	1	1	0
27	1	68	1.26	1	1	1							

are a few interesting things to note about the values of m_2 that one obtains. The first is that they are relatively robust to the prior specification(except for model 2) which suggests that the prior used provides only limited information over the flat prior. Moreover, the ordering of the models is roughly the same as that obtained from the

Table 2.2
Replication Design Matrix

X_1	X_2	X_3	X_4	X_5
56	0.47	1	1	1
62	0.71	1	1	0
58	0.57	1	0	1
59	0.65	1	0	0
61	0.69	0	1	1
63	0.64	0	1	0
59	1.06	0	0	1
58	0.85	0	0	0

marginal, though the amount of separation between various models decreases when using m_2.

Table 2.3
Model Comparison for Nodal Data

		Proper Prior		Flat Prior
Model	$\log\left(f(\boldsymbol{y}\|\widehat{\boldsymbol{\beta}}_{\boldsymbol{y}})\right)$	$\log\left(m(\boldsymbol{y})\right)$	$\log\left(m_2(\boldsymbol{y})\right)$	$\log\left(m_2(\boldsymbol{y})\right)$
C	-35.1	-38.5	-36.3	-36.3
$C + x_1$	-34.6	-43.2	-36.9	-37.8
$C + x_2$	-32.4	-37.9	-34.9	-34.9
$C + x_3$	-29.5	-35.3	-31.6	-31.6
$C + x_4$	-31.3	-37.2	-33.5	-33.5
$C + x_5$	-33.1	-39.1	-35.3	-35.3
$C + x_2 + x_4$	-28.2	-36.1	-31.7	-31.8
$C + x_2 + x_3 + x_4$	-24.4	-34.5	-28.9	-29.1
$C + x_2 + x_3 + x_4 + x_5$	-23.8	-36.2	-29.2	-29.5

Example 2.3.9 (Balanced Random Effects Model)**.** *Another model where the utility of the* $\alpha = 2$ *case is easily seen is the balanced random effects model. This model is particularly easy to analyze because the posterior distribution can be readily sampled*

from if one uses an appropriate improper prior. We assume that $y_{ij} = \mu_j + \epsilon_{ij}$ *and* $\mu_j = \mu + u_j$ *for unit* i *in group* j *for* $i = 1, \ldots, n$ *and* $j = 1, \ldots, J$ *where* $\epsilon_{ij} \sim \mathcal{N}(0, \tau^{-1})$ *and* $u_j \sim \mathcal{N}(0, (\tau r)^{-1})$. *We analyze this model using the independence Jeffreys' prior,* $\pi(\mu, \tau, r) = \frac{cn}{\tau r(n+r)}$. *Integrating out the* u_j *yields a set of vectors* $\boldsymbol{y}_j = (y_{1j}, \ldots, y_{nj})'$ *for* $j = 1, \ldots, J$ *where* $\boldsymbol{y}_j \sim \mathcal{N}(\mathbf{1}_n \mu, \tau^{-1}(\boldsymbol{I}_n + r^{-1}\mathbf{1}_n\mathbf{1}_n'))$. *In this formulation of the model,* $r = 0$ *corresponds to the fixed effects model and* $r = \infty$ *corresponds to* $u_j = 0 \ \forall j$.

For convenience, define $\boldsymbol{y} = (\boldsymbol{y}_1', \ldots, \boldsymbol{y}_J')'$ *and*

$$SSB = n\sum_{j=1}^{J}(\overline{\boldsymbol{y}_j}^2 - \overline{\boldsymbol{y}}^2), \quad SSW = \sum_{j=1}^{J}(\boldsymbol{y}_j'\boldsymbol{y}_j - n\overline{\boldsymbol{y}_j}^2), \quad SST = SSB + SSW$$

The marginal density under this prior depends only on the summary statistics SSB *and* SSW *as well as the constant* c. *In fact, using the change of variables* $t = \frac{SSB}{SSW}\frac{r}{n+r}$ *provides*

$$m(\boldsymbol{y}) = \frac{c\Gamma\left(\frac{nJ-1}{2}\right)}{J^{\frac{1}{2}}\pi^{\frac{nJ-1}{2}}SSB^{\frac{J-1}{2}}SSW^{\frac{(n-1)J}{2}}}B\left(\frac{SSB}{SSW}; \frac{J-1}{2}, \frac{(n-1)J}{2}\right)$$

where $B(x; a, b) = \int_0^x \frac{t^{a-1}}{(1+t)^{(a+b)}}dx$ *is the incomplete beta function.*

The posterior distribution is expressed through conditional distributions as

$$\begin{aligned}
\mu|\tau, r, \boldsymbol{y} &\sim \mathcal{N}\left(\overline{\boldsymbol{y}}, \frac{n+r}{\tau r n J}\right) \\
\tau|r, \boldsymbol{y} &\sim \textit{Gamma}\left(\frac{nJ-1}{2}, \frac{rSST + nSSW}{2(n+r)}\right) \\
\frac{rSSB}{rSST + nSSW}|\boldsymbol{y} &\sim \textit{Truncated Beta}\left(\frac{J-1}{2}, \frac{(n-1)J}{2}; \frac{SSB}{SST}\right)
\end{aligned}$$

The posterior predictive density of a sample $\boldsymbol{z}$ *with* J' *groups is* $\int p(\boldsymbol{z}|r,\boldsymbol{y})p(r|\boldsymbol{y})dr$ *where*

$$\boldsymbol{z}|r,\boldsymbol{y} \sim \mathcal{MVT}_{nJ'}\left(nJ-1,\overline{\boldsymbol{y}},\Sigma_r\right)$$

and it is assumed that $\boldsymbol{z} \perp\!\!\!\perp \boldsymbol{y}|\mu,\tau,r$. *The covariance matrix is*

$$\Sigma_r = \left(\frac{SSW+\frac{r}{n+r}SSB}{nJ-1}\right)\left(\boldsymbol{I}_{nJ'}+\frac{1}{r}\boldsymbol{I}_{J'}\otimes(\mathbf{1}_n\mathbf{1}_n')+\frac{n+r}{nJr}(\mathbf{1}_{nJ'}\mathbf{1}_{nJ'}')\right)$$

where $\otimes$ *is the Kronecker product.*

Since the integral in $\boldsymbol{z}$ *cannot be done analytically, a simple MCMC algorithm must be implemented to compute the criterion. First, we generate random variates* $r^{(1)},\ldots,r^{(N)}$ *using the truncated beta distribution. Then, for each* $r^{(i)}$ *generate a* $\boldsymbol{z}^{(i)}$ *using the multivariate t-distribution. Finally, compute*

$$\left(\frac{1}{N}\sum_{i=1}^{N}\left(\frac{m(\boldsymbol{z}^{(i)},\boldsymbol{y})}{m(\boldsymbol{z}^{(i)})}\right)^{\alpha-1}\right)^{\frac{1}{\alpha-1}}$$

using the formulas for the marginal distribution. The marginal for $(\boldsymbol{y},\boldsymbol{z})$ *can be easily computed using the following formulas for sums of squared errors*

$$SSW_{(\boldsymbol{y},\boldsymbol{z})} = SSW_{\boldsymbol{y}}+SSW_{\boldsymbol{z}}, \quad SSB_{(\boldsymbol{y},\boldsymbol{z})} = SSB_{\boldsymbol{y}}+SSB_{\boldsymbol{z}}+\frac{nJJ'}{J+J'}(\overline{\boldsymbol{y}}-\overline{\boldsymbol{y}})^2.$$

We compare the ratio of modified marginals between random effects model and the constant mean model to the many Bayes Factors obtained in [50] for the [51] dyestuff data (both actual and simulated). The constant mean model was analyzed using the reference prior $\pi(\mu,\tau) \propto \frac{1}{\tau}$. *Provided in 2.3.9is a table of values for the modified marginal for various values of* α *as well as marginals obtained through various*

methods (see [50] for a discussion). In addition, we also present the median intrinsic Bayes' Factor using partial Bayes Factors' which come from samples comprising of two groups.

Table 2.4
Model Comparison for Dyestuff Data

Method	Actual Dyestuff	Simulated Dyestuff
WG	11.85	0.17
DB, $q = 0.75$	5.25	0.11
DB, $q = 0.1$	7.88	0.18
DB, $q = 1.25$	9.35	0.23
I^*	8.90	0.16
MI	4.68	0.11
$\overline{B}_{21}$	15.4	1
MoBF, $\alpha = 0.5$	1.77	0.06
MoBF, $\alpha = 1.5$	3.82	0.12
MoBF, $\alpha = 2$	4.23	0.14
$MIBF^*$	7.40	0.30

This comparison highlights a fact previously seen with the Haldane prior example: integrating out the uncertainty in $\boldsymbol{z}$ often makes this criterion more conservative than existing methods. For the dyestuff data, the postive evidence in favor of the random effects model using $\alpha = 2$ is 4.22 using this approach, which is more conservative than the minimum (4.68) of those presented by [50]. For the simulated data from [51], the strong evidence against the random effects model is 0.14, which is more conservative than 4 of the 6 Bayes factors obtained in [50]. It is easy to show the the modified Bayes' Factor produces consistent model selection when n is fixed and $J \to \infty$ (arguably the point of such a model), however, it is not known whether these

modified Bayes' Factors will produce consistency when $n \to \infty$ for a fixed number of groups J. However, if one uses a minimal sample which consists of two groups each with two observations (regardless of the size of the observations in the actual data), it is easy to see that consistency is achieved across the class of models.

2.4 Computational Issues

Though these criteria offer theoretical advantage over existing methods of model selection in the case when one has vague prior knowledge, they can be computationally burdensome. In particular, each method requires a number of integrations. In the PPIC case, all of these integrations can be computed using MCMC methods, but one then needs to take a double sum, one over the parameter space in order to compute the predictive densities and one over the space of replicated samples. There are two ways to speed up this computational process. The first is that one can parallelize the estimation of the predictive densities after the samples over the parameter space has been taken. the other is to produce only one sample of replicated data and one use importance sampling. Importance sampling arises from the fact that

$$I = \int \phi(x) f(x) \mathrm{d}x = \int \phi(x) \frac{f(x)}{g(x)} g(x) \mathrm{d}(x)$$

to change the estimator $I \approx \frac{1}{N} \sum \phi(x^{(i)})$ where samples are taken from data generated by f and replaces it with the estimator $I \approx \frac{1}{N} \sum \phi(x^{(i)}) \frac{f(x^{(i)})}{g(x^{(i)})}$ where samples are taken from data generated by f. Using importance sampling, samples can be taken from

only one predictive density in order to compute all of the integrals in the signed divergences.

Computational issues also arise from the modified modified marginals, but they arise in two ways. First, one must compute marginal distributions, which can be computationally burdensome if one cannot integrate them analytically or with a quick numerical method (such as quadrature. If one has to compute the marginals from MCMC samples, there are essentially two choices for overcoming this problem. The first is to notice that

$$\frac{1}{m(x)} = \int \frac{\pi(\theta)}{m(x)} \mathrm{d}x = \int \frac{f(\theta|x)}{f(x|\theta)} \mathrm{d}x$$

when one has access to a proper prior, allowing one to use a harmonic estimator:

$$m(x) = \left(\frac{1}{N} \sum_i \frac{1}{f(x|\theta(i))} \right)^{-1}.$$

However, the merit of the criterion is that it can be used when proper priors are not available, and so this method might create some nonsensical estimators. Moreover, this method is computationally unstable. A correction to this method is available from Gelfand and Dey [21], but it requires a nice tuning function. The best way to overcome this difficulty is to use the method of Chib [49, 52], where the estimator is replaced by

$$\widehat{m}(x) = \frac{f(x|\theta^*)\pi(\theta^*)}{p(\theta^*|x)}$$

where $p(\theta^*|x)$ is computed using the availability of the closed form posterior conditioned on both x and some "missing" data, integrating out the missing data by

summing over samples of it and θ^* is an appropriately chosen ordinate. This is the best method when exact computations are not available, and requires the additional sampling of the missing data for each sample from the predictive density, which adds an additional computational burden. There is yet another computational issue when one is computing the modified marginal for fractional α, which is that the sample average is itself a harmonic mean, and so exhibits the same instability as does computing a marginal using the harmonic mean. Essentially, sets of small probability have a large impact on the Rényi entropies when $0 < \alpha < 1$ and so a few samples that arise with low probability have a large impact on the estimated criterion.

2.5 Conclusions

We have presented two new methods of model comparison and selection, one which is model external and decomposes as fit and complexity terms and another that maintains the status of the Bayes' Factor as the primary tool of model comparison, even for improper priors. Both methods allow for the investigator to tune them in order to achieve some inferential goal. In the case of the PPIC, α can be set to control the Type I error rate or modified in such a way to get asymptotic consistency. In the case of the modified marginals, the investigator gets to choose both the design of the replicated data $\boldsymbol{z}$ and the particular Rényi divergence used. In this manner, the investigator can control the Type I error rate in small samples while still maintaining the asymptotic equivalence to the Bayes' Factor.

There are open questions which remain from both a theoretical and computational perspective. From a theoretical perspective, we would like to determine if for fixed $0 < \alpha \leq 1$ and a minimal sample what value the Rényi divergences take. Also, there is no generic proof about the behavior of the criteria if one uses fully replicated data and allows $\alpha \to 0$ in an appropriate manner. For the PPIC, there is no generic way to determine how one needs to allow α to vary in order to obtain consistency and at this point it needs to be determined for each problem. Additionally, the extension to multiple models required the existence of a number of information integrals. If one of these integrals happens to be infinite, what can one do in order to repair the method. One possible solution is to remove the appropriate term from each summand, but this seems like an ad-hoc way to remedy the issue. Computationally, one needs to find ways to quickly compute all of the integrals that need to be obtained. Though there are ways to address these issues, each adds a layer of sampling and summation, which becomes computationally taxing as the number of replicated samples becomes large.

REFERENCES

[1] A.N. Kolmogorov. *Foundations of the theory of probability.* Chelsea Publishing Co., 1950.

[2] P. Billingsley. *Probability and Measure.* 1995.

[3] A. Rényi. Sur les espaces simples de probabilités conditionelles. 1:3–21, 1964.

[4] R.T. Cox. *The algebra of probable inference.* Johns Hopkins Press, 1961.

[5] E.T. Jaynes and G.L. Bretthorst. *Probability theory: the logic of science.* Cambridge Univ Pr, 2003.

[6] J.Y. Halpern. Technical Addendum Cox's Theorem Revisited. *Journal of Artificial Intelligence Research*, 11(429):435, 1999.

[7] E. Hewitt and L.J. Savage. Symmetric measures on Cartesian products. *Trans. Amer. Math. Soc*, 80(1):470–501, 1955.

[8] P. Diaconis and D. Freedman. Finite exchangeable sequences. *The Annals of Probability*, 8(4):745–764, 1980.

[9] J.K. Ghosh and RV Ramamoorthi. *Bayesian nonparametrics*. Springer Verlag, 2003.

[10] L.J. Savage. *The foundations of statistics*. Dover Pubns, 1972.

[11] H. Jeffreys. *Theory of probability*. Oxford University Press, USA, 1998.

[12] J. Berger. The case for objective Bayesian analysis. *Bayesian Analysis*, 1(3):385–402, 2006.

[13] J. Aldrich. RA Fisher on Bayes and Bayes theorem. *Bayesian Analysis*, 3(1):161–170, 2008.

[14] G.S. Datta and J.K. Ghosh. Noninformative priors for maximal invariant parameter in group models. *Test*, 4(1):95–114, 1995.

[15] H. Jeffreys. An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 186(1007):453–461, 1946.

[16] J.M. Bernardo. Reference posterior distributions for Bayesian inference. *Journal of the Royal Statistical Society. Series B (Methodological)*, 41(2):113–147, 1979.

[17] J.O. Berger and J.M. Bernardo. On the development of reference priors. *Bayesian statistics*, 4:35–60, 1992.

[18] D.S. Sivia and J. Skilling. *Data analysis: a Bayesian tutorial*. Oxford University Press, USA, 2006.

[19] R.E. Kass and A.E. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90(430), 1995.

[20] G. Schwarz. Estimating the dimension of a model. *The annals of statistics*, pages 461–464, 1978.

[21] A.E. Gelfand and D.K. Dey. Bayesian model choice: asymptotics and exact calculations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 56(3):501–514, 1994.

[22] D.V. Lindley. A statistical paradox. *Biometrika*, 44(1/2):187–192, 1957.

[23] H. Akaike. A Bayesian analysis of the minimum AIC procedure. *Annals of the Institute of Statistical mathematics*, 30(1):9–14, 1978.

[24] H. Bozdogan. Akaike's information criterion and recent developments in information complexity. *Journal of mathematical psychology*, 44(1):62–91, 2000.

[25] M. Aitkin. Posterior Bayes factors. *Journal of the Royal Statistical Society. Series B (Methodological)*, 53(1):111–142, 1991.

[26] S. Konishi and G. Kitagawa. Generalised information criteria in model selection. *Biometrika*, 83(4):875, 1996.

[27] D.J. Spiegelhalter, N.G. Best, B.P. Carlin, and A. van der Linde. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, 64(4):583–639, 2002.

[28] A.E. Gelfand and S.K. Ghosh. Model choice: a minimum posterior predictive loss approach. *Biometrika*, 85(1):1, 1998.

[29] S. Geisser and W.F. Eddy. A predictive approach to model selection. *Journal of the American Statistical Association*, 74(365):153–160, 1979.

[30] J.O. Berger and L.R. Pericchi. The intrinsic Bayes factor for model selection and prediction. *Journal of the American Statistical Association*, 91(433):109–122, 1996.

[31] E. Moreno, F. Bertolino, and W. Racugno. An Intrinsic Limiting Procedure for Model Selection and Hypotheses Testing. *Journal of the American Statistical Association*, 93(444):1451–1452, 1998.

[32] G. Casella, F.J. Girón, M.L. Martınez, and E. Moreno. Consistency of Bayesian procedures for variable selection. *The Annals of Statistics*, 37(3):1207–1228, 2009.

[33] G. Casella and E. Moreno. Objective Bayesian variable selection. *Journal of the American Statistical Association*, 101(473):157–167, 2006.

[34] E. Moreno, G. Casella, and A. Garcia-Ferrer. An objective Bayesian analysis of the change point problem. *Stochastic Environmental Research and Risk Assessment*, 19(3):191–204, 2005.

[35] A. O'Hagan. Fractional Bayes factors for model comparison. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):99–138, 1995.

[36] A. OHagan. Properties of intrinsic and fractional Bayes factors. *Test*, 6(1):101–118, 1997.

[37] J.M. Pérez and J.O. Berger. Expected-posterior prior distributions for model selection. *Biometrika*, 89(3):491–511, 2002.

[38] K. Iwaki. Posterior expected marginal likelihood for testing hypotheses. *J. Econ. Asia Univ*, 21:105–34, 1997.

[39] C. Shannon. The mathematical theory of communication. *The Bell System Technical Journal*, 27(7):379–423, 1948.

[40] I. Csiszar. I-Divergence Geometry of Probability Distributions and Minimization Problems. *The Annals of Probability*, 3(1):146–158, 1975.

[41] I. Csiszár and P.C. Shields. *Information theory and statistics: A tutorial*. Now Publishers Inc, 2004.

[42] S.I. Amari. Differential geometry of curved exponential families-curvatures and information loss. *The Annals of Statistics*, 10(2):357–385, 1982.

[43] S. Amari. *alpha*-Divergence Is Unique, Belonging to Both f-Divergence and Bregman Divergence Classes. *Information Theory, IEEE Transactions on*, 55(11):4925–4931, 2009.

[44] P.W. Vos. Geometry of f-divergence. *Annals of the Institute of Statistical Mathematics*, 43(3):515–537, 1991.

[45] A. Rényi. On measures of entropy and information. In *Fourth Berkeley Symposium on Mathematical Statistics and Probability*, pages 547–561, 1961.

[46] S. Kullback and RA Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, pages 79–86, 1951.

[47] L. Leon-Novelo, E. Moreno, and G. Casella. Objective Bayes Model Selection in Probit Models. 2011.

[48] J.H. Albert and S. Chib. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88(422):669–679, 1993.

[49] S. Chib. Marginal Likelihood from the Gibbs Output. *Journal of the American Statistical Association*, 90(432), 1995.

[50] G. García-Donato and D. Sun. Objective priors for hypothesis testing in one-way random effects models. *Canadian Journal of Statistics-Revue Canadienne de Statistique*, 35(2):303–320, 2007.

[51] G.E. Box and G.C. Tiao. Bayesian inference in statistical analysis. 1973.

[52] S. Chib and I. Jeliazkov. Marginal likelihood from the Metropolis-Hastings output. *Journal of the American Statistical Association*, 96(453):270–281, 2001.